Utilisation of novel temporal and traditional GWAS strategies to elucidate key factors contributing to the success of *Shigella* species as pathogens

Thesis submitted in accordance with the requirements of the University of Liverpool for the degree of Doctor in Philosophy by

Rebecca Jane Bennett

September 2022

Abstract

Title of thesis: Utilisation of novel temporal and traditional GWAS strategies to elucidate key factors contributing to the success of Shigella species as pathogens.

Author: Rebecca Jane Bennett

Bacteria of the genus *Shigella* are a major contributor to the global diarrhoea burden causing >200,000 deaths per annum globally. Increasing antibiotic resistance in *Shigella* and the lack of a licenced vaccine has led WHO to recognise *Shigella* as a priority organism for the development of new antibiotics. Understanding what drives the long-term success of this pathogen is critical for ongoing global management of shigellosis and has relevance for other enteric bacteria.

To identify key genetic drivers of *Shigella* evolution over the past 100 years, the unique and significant potential of historical bacterial genomes was utilised. The historical Murray collection, comprising several hundred pre-antibiotic era (1917 – 1954) *Enterobacteriaceae,* was used alongside comparatively modern (1950s – 2018) isolates to conduct GWAS within both *S. flexneri* and *S. sonnei*. Within *S. flexneri* I identified 94 SNPs and significant kmers within 48 genes significantly positively associated with time as a continuous variable. These included genes encoding T3SS proteins, proteins involved in intracellular competition, and multi drug resistance proteins which have intuitively beneficial roles for *Shigella* as pathogens. However, 34% of identified hits related to genes of as yet unknown function. Subsequent Alphafold modelling of on such protein has identified one of these hits as a putative novel adhesin, adhesion Stv. Stv has been shown to be conserved through PG expansion in PG 1 and PG 6 in *S. flexneri* highlighting its key potential role in successful *S. flexneri* infection. Furthermore, presence of this adhesin within *S. sonnei* revealed its potential contribution to the global success of Lineage III making this an exciting validation of the adhesin as a factor contributing to the success of *Shigella* species over time.

Thus, my temporal GWAS approach, has identified known and novel genetic factors that are enriching in *Shigella* populations over time. These genetic determinants may thus be key factors in the long-term success of *Shigella* as well as offer potential targets for pathogen management. Similarly in *S. sonnei* temporal GWAS has revealed exciting factors which are intuitively beneficial for the success of *S. sonnei* over time, supporting the robustness of this novel methodology.

In addition to temporal GWAS, traditional GWAS on a collection of Lineage III *S. sonnei* was completed to elucidate what factors were behind their advantageous E. coli killing ability. GWAS revealed colicins, small proteins which are toxic to *E. coli*, as the potential factors resulting in this phenotype. Further screening of the isolates for potential colicins within the isolates revealed that it was not a singular colicin responsible but a wide array of colicins contributing to the phenotype.

Through both temporal GWAS and categorical GWAS, novel factors which are key to the success of *Shigella* species have been identified.

Acknowledgements

This PhD has simultaneously been the best experience of my life whilst also feeling like I was trying to solve a 10,000-piece jigsaw puzzle without a box. And the pieces keep changing shape and colour. And the room is on fire. But genuinely I could not have completed my PhD without the help and support of so many people.

Firstly, I would like to thank my primary supervisor, Kate Baker, for her continuous support and guidance throughout my time in the Bakery. I would additionally like to thank my secondary supervisor, Mal Horsburgh, who welcomed me into the lab when I was just a first year undergraduate and gave me the initial confidence to pursue a PhD. You opened my eyes to research, and I stayed so long thanks to you. Finally, from the supervisor team I would like to thank my other secondary supervisor, Tim Blower. Your enthusiasm and kindness are so wonderful to have as a supervisor. With your training and guidance my PhD has gone from strength to strength.

I have been fortunate enough to have been part of the great community that is Lab H and I have met so many people that made my time as a PhD student so fun and worth every second of stress. Specifically, I would like to thank the Bakery; Rebecca Bengtsson, P. Malaka De Silva, George Stenhouse, Charlotte Chong and Lewis Mason. Thank you for listening to my ideas, solving my problems, and making the Bakery a great group to work with. I am particularly indebted to P. Malaka De Silva who is our saviour in the lab! Thank you for helping me with adhesin Stv and our joint colicin paper is a great demonstration of how lab and bioinformatic work so well together. I would also like to thank other members of Lab H; Jack Fitzpatrick, Ross Mulhall, Ella Rodwell, Caisey Pulford, Rama Bhatia, and Hermione Webster. Thank you for making Lab H a great place to work.

My PhD could also not have been possible without my amazing funders from the BBSRC and our collaborators from around the globe. BBSRC gave me the funding to pursue this amazing project and have provided support, training, and great conferences every year of my PhD. I would also like to thank the team from UKHSA, particularly Sarah Alexander and the team from NCTC who curated the historical isolates which were so vital to my project.

Finally, I would like to thank my incredibly supportive family. I feel honoured and blessed to have such an amazing family who have always provided me with confidence, encouragement, and support. I am always grateful to my Mum, for her unconditional love and who has always listened to me when I am stressed and given me the belief to carry one. And to my brother who always makes me laugh and tries to help even when he has no idea what I am doing (normally with a glass of wine). You have taught me how to write and maybe use a comma correct occasionally. Although my Dad did not make it to the finish line with me, I hope he's with me in spirit. Thank you for giving me the best life advice – "Always read the question" and "Never turn down a burrito". You are deeply missed.

This thesis is dedicated to my Dad,
Steven John Bennett,
who believed that I could do anything.

Publications

Below is a list of publications written throughout the course of this PhD. Publications where I am a first or joint-first author are indicated with an asterisk.

Publications in peer-reviewed journals

Bennett RJ*, Baker KS. Looking Backward To Move Forward: the Utility of Sequencing Historical Bacterial Genomes. J Clin Microbiol. 2019 Jul 26;57(8):e00100-19. doi: 10.1128/JCM.00100-19. PMID: 31092597; PMCID: PMC6663899.

Rebecca J. Bengtsson, Timothy J. Dallman, Hester Allen, P. Malaka De Silva, George Stenhouse, Caisey V. Pulford, **Rebecca J. Bennett**, Claire Jenkins, Kate S. Baker Accessory Genome Dynamics and Structural Variation of Shigella from Persistent Infections DOI: https://doi.org/10.1128/mBio.00254-21

Caisey V. Pulford, Blanca M. Perez-Sepulveda, Rocío Canals, Jessica A. Bevington, Rebecca J. Bengtsson, Nicolas Wenner, Ella V. Rodwell, Benjamin Kumwenda, Xiaojun Zhu, **Rebecca J. Bennett,** George E. Stenhouse, P. Malaka De Silva, Hermione J. Webster, Jose A. Bengoechea, Amy Dumigan, Alicia Tran-Dien, Reenesh Prakash, Happy C. Banda, Lovemore Alufandika, Mike P. Mautanga, Arthur Bowers-Barnard, Alexandra Y. Beliavskaia, Alexander V. Predeus, Will P. M. Rowe, Alistair C. Darby, Neil Hall, François-Xavier Weill, Melita A. Gordon, Nicholas A. Feasey, Kate S. Baker & Jay C. D. Hinton. Stepwise evolution of Salmonella Typhimurium ST313 causing bloodstream infection in Africa. Nat Microbiol **6**, 327–338 (2021). https://doi.org/10.1038/s41564-020-00836-1

Publications under review in peer-reviewed journals

Rebecca J Bennett*, P. Malaka De Silva, Rebecca J Bengtsson, Malcolm J Horsburgh, Tim R Blower, Kate S Baker Temporal GWAS identifies a widely distributed putative adhesin contributing to pathogen success in Shigella spp bioRxiv 2022.08.23.504947; doi: https://doi.org/10.1101/2022.08.23.504947

P. Malaka De Silva, **Rebecca J. Bennett***, Lauriane Kuhn, Patryk Ngondo, Brian Ho, Francois-Xavier Weill, Benoît S. Marteyn, Claire Jenkins, Kate S. Baker *Escherichia coli* killing by *Shigella sonnei* is mediated by colicins found in epidemiologically successful sublineages

Pre-Contents

Table of Contents

List of Figures

List of Tables

Abbreviations

ACME	Arginine Catabolic Mobile Element
aDNA	Ancient DNA
AI	Artificial intelligence
AMR	Antimicrobial resistance
ARG	Antimicrobial resistance gene
ATCC	American Type Culture Collection
BWA	Burrows wheeler aligner
CARD	Comprehensive Antibiotic Resistance Database
CF	Cystic fibrosis
CIP	The Institut Pasteur Collection
COG	Clusters of orthologous Groups
ENA	European nucleotide archive
GEMS	Global Enteric Multicenter Study
GWAS	Genome Wide Association study
HGT	Horizontal Gene Transfer
IBC	Intracellular bacterial communities
IS	Insertion sequences
Kb	Kilobases
LMIC	Lower to middle income countries
LMM	Linear mixed model
LPS	Lipopolysaccharide
MDR	Multi Drug Resistant
MGE	Mobile Genetic Element
MRCA	Most recent common ancestor
MRSA	Methicillin resistant Staphylococcus aureus
MSM	Men who have sex with men
NCBI	National Center for Biotechnological Information
NCTC	National Collection of type cultures
NGS	Next generation sequencing
OJC	Orthodox Jewish Communities
ONT	Oxford Nanopore technologies
PAI	Pathogenicity island
PAMP	Pathogen-associated molecular pattern
PG	Phylogroups
PGAP	Prokaryotic Genome Annotation Pipeline
PGWAS	Pangenome genome wide association study
pINV	The virulence plasmid for *Shigella*
SM	Specialised metabolites
SMRT	Single-molecule real-time sequencing
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
STI	Sexually transmitted infection
T3SS	Type III Secretion system
T6SS	Type VI Secretion system

TA	Toxin: antitoxin
tGWAS	Temporal Genome wide association study
UKHSA	UK Health and Security Agency
VFDB	Virulence finder database
VRSA	Vancomycin resistant Staphylococcus aureus
WHO	World health organisation

Chapter 1

1. Introduction

This PhD project is a true amalgamation of historical pathogens and modern biological fields of research. The evolution of bacteria is not passive but is guided by selective pressures caused by human events such as war, implementation of antibiotics and human migrations. To fully understand evolution of public health pathogens it is important to conduct investigations incorporating isolates spanning longer time periods. This allows us to understand how human events have shaped bacteria and in turn better predict how bacteria will be shaped in the future allowing improved prevention, treatment, and management.

Some of the content within this chapter, specifically section 1.2, was presented in the review article 'Looking Backward To Move Forward: the Utility of Sequencing Historical Bacterial Genomes' published in the Journal of Clinical Microbiology by the American Society for Microbiology. Permission to include the publication in this PhD thesis was obtained from all co-authors.

1.1. The long-term success of bacterial species

There are a plethora of factors and selective pressures behind pathogen success and their continued persistence globally. The following sections first aim to explore historical pathogens which still persist today followed by traditional methodologies for researching their evolutionary arcs and finally exploring the factors which have led to the pathogenic success of these persistent pathogens over time.

1.1.1. The secret war: Bacteria versus humans

Since the beginning there has been a secret arms race for survival between humans and pathogens. The following sections aim to explore and give a brief overview of some of the historical pathogens which still continue to cause impactful disease and burden globally followed by some of the ways bacteria have evolved in response to human interventions and immunity to aid pathogen success.

1.1.1.1. Long term persistence of public health pathogens

In 2022 the human population is no stranger to the extensive and devastating impacts which can occur from the emergence of an infectious disease. COVID-19 has highlighted the importance of studies into the evolution, prevention, and treatment of microorganisms. For centuries, pandemics, and major epidemics such as plague, cholera and flu have left their devastating mark on humanity. Furthermore, many of the bacterial species that were the causative agents for these historical pandemics are still causing widespread infections today. Here long-term persistence and success is defined as a historical pathogen which has continued to cause successful and impactful disease in the modern era.

For example, the bacterium *Vibrio cholerae* is the causative agent for cholera, an acute gastrointestinal disease characterised throughout history by the rapid and massive loss of fluids leading to severe dehydration and possible fatality (Deen et al., 2020). By the 19th century this prominent pathogen had caused seven global pandemics with the seventh pandemic still ongoing today (Bennett and Baker, 2019). The seventh pandemic is thought to still be infecting 3-5 million people annually (Hu et al., 2016). This is a prime example of the persistence of a bacterial species within the human population.

The long-term persistence of bacterial species can also clearly be seen in the family of *Enterobacteriaceae* for example in *Escherichia coli* and, the focus of this thesis, *Shigella*. The natural history of *Shigella* is a lengthy and enduring one. *Shigella*, the causative agent of

bacillary dysentery, was first identified in 1898 by Kiyoshi Shiga but the idea of dysentery has been recorded since at least the time of Hippocrates (Brenner et al., 2005, Davison, 1922). In 1987, an outbreak of bacillary dysentery in Japan killed over 22,000 people with a mortality rate of 25% and today *Shigella* species cause approximately 216,000 deaths annually (Lampel et al., 2018, Khalil et al., 2018). The bacteria of the *Enterobacteriaceae* family are globally important and the long-term persistence of these species poses a huge burden to global public health.

1.1.1.2. The burden of bacterial disease

The long-term persistence of historical pathogens often leads to a heavy burden on public health including increased mortality, morbidity, and financial expense. Even with the rapid advancement in medical interventions, such as antibiotics and vaccines, these bacterial infections contribute heavily to the global burden on healthcare systems owing to their ability to outstrip our medical advancements through continued evolution.

For example, the global burden of cholera is currently thought to be vastly underestimated, mainly due to poor surveillance and low reporting in low- to middle-income countries (LMIC) (Fournier and Quilici, 2007). The World Health Organisation (WHO) stated that only 2000-3000 deaths are officially reported to them each year. However, the actual mortality numbers are estimated to be over 95,000 deaths and 2.9 million cases annually (Fournier and Quilici, 2007, Ilboudo et al., 2017). With the large quantity of global cases, it comes with a heavy economic burden contributed to by treatments and extended hospital stays. It is estimated that the cost of cholera related illness and treatment in Africa during 2015 was US$130 million (Mogasale et al., 2021).

It is evident that historical pathogens are still causing significant numbers of infections and contributing to the global burden of bacterial infections on healthcare. Although ingenious

advancements in treatments and preventative measures have been effective in curbing infection numbers and mortality rates, historical pathogens are still of global concern. Understanding how these pathogens evolve and adapt to persist globally could help to reduce the rates of infections from these pathogens.

1.1.1.3. The evolution of bacterial warfare

In response to advancements in treatments and preventative measures against pathogens, bacteria are continually evolving a wide range of mechanisms to evade and resist these advancements. The introduction of novel antibiotics throughout the past 50 years has contributed to the positive selection for acquisition of resistance determinants and so requires the implementation of new treatment regimens to "keep up". The selective pressure caused by both the immune system and human interventions positively select for the evolution of bacterial virulence within public health pathogens. This evolution of bacterial warfare is key to the prolonged success of pathogens.

There are many ways in which bacteria have evolved to remain successful as pathogens even with the ever more sophisticated human interventions, bacteria still manage to in some cases to have the upper hand. Below are some of the key areas which I believe deserve further discussion.

1.1.1.3.1. The increase in AMR in response to human interventions

One dominant area of public health concern is antimicrobial resistance (AMR). The WHO has warned of a "post antibiotic" era where even the simplest of infections could kill (Region et al.). In 2019 it had been estimated that there were 4.95 million deaths associated with bacterial with AMR and a recent report claimed there would be 10 million deaths by 2050 (Murray et al., 2022, O'Neil, 2014). There is a need for the discovery of novel antibiotics or

alternative treatments as well as a deeper understanding of AMR evolution in prominent bacterial species.

Since the discovery of penicillin by Alexander Fleming in 1928 there has been increased use and misuse of a wide variety of antibiotics. The first report of resistance to penicillin was in 1947 within the Gram-positive bacterium *Staphylococcus aureus* which has evolved to become a prominent "superbug" in the modern day (Podolsky, 2018). Through the evolution of de novo mutations and the acquisition of horizontally transferred resistance determinants such as plasmids, *S. aureus* has evolved resistance to multiple antibiotic resistant determinants including methicillin in 1961 (Hiramatsu, 2001). Methicillin-resistant *S. aureus* (MRSA) in 2019 was the second leading cause of deaths associated with MDR bacteria and caused an estimated 600,000 deaths (Murray et al., 2022). There has been another worrying development in *S. aureus*, namely resistance to the glycopeptide vancomycin, often considered the last line therapeutic agent for the treatment of MRSA infections and other Gram-positive infections (Wu et al., 2021). Since the first reported case of vancomycin-resistant *S .aureus (*VRSA) in 2001 there has been a 3.5 fold increase in the frequency of VRSA infections from 2006 to 2020 putting tremendous pressure on public health systems to find alternative effective treatment (Wu et al., 2021).

The acquisition of AMR determinants has been well documented as a contributor to the prolonged success of bacterial pathogens. Through different mechanisms bacteria are able to acquire genes, plasmids and mutations which confer AMR resistance to a wide variety of antibiotics leading to a lack of efective treatment options.

Although AMR is a major contributor to pathogen success and widely studied in public health, there are a plethora of other factors which all contribute to overall success and have evolved due to human hosts. This is evident from USA300, a very successful MRSA strain

first identified in the US, where although AMR was key, other key elements were identified unique to these strains which contribute profoundly to its global success. Some of these other potential factors will be discussed further below.

1.1.1.3.2. Bacteria's evolution to evade the host immune system

It is not only through treatment options that bacteria have evolved to resist or evade. Prominent public health pathogens all impose a heavy selection pressure on human hosts to evolve resistance mechanisms to disease. The innate and adaptive immune systems have evolved under the selective pressure of microorganisms since their introduction (Reddick and Alto, 2014, Flajnik and Kasahara, 2010). Today in modern eukaryotic organisms the immune mechanisms are highly sophisticated, complex, and effective in combating infection involving multiple components and signalling cascades (Reddick and Alto, 2014). Conversely in response to the evolution of a highly effective immune system, bacterial pathogens have developed their own mechanisms to evade and inhibit host immune systems.

Typically, the initiation of the immune system involves the recognition of the external surface of a bacterial pathogen for immune surveillance. However, this also provides an opportunity for bacteria to develop mechanisms to mimic, inhibit or alter immune system components. A common mechanism for bacteria to "hide" from the immune system is through expression of a carbohydrate capsule which functions to shield the complex bacterial surface of proteins and carbohydrates from immune surveillance (Finlay and McFadden, 2006). The capsule mechanism is utilised in many bacteria including in *Streptococcus pneumoniae* where it has been well established as a major virulence factor to evade phagocytosis (Sanders et al., 2011).

Most host organisms have evolved a defence system reliant on detecting pathogen-associated molecular (PAMPs) which are small molecular motifs conserved across multiple bacterial species and recognised by toll-like receptors. LPS, a glycolipid located on the outside of Gram-negative bacteria, is detected as a PAMP, specifically its lipid A component (Paciello et al., 2013). The outer part of the LPS is highly variable and historical pathogens have frequently exploited this. *Yersinia pestis*, the causative agent of the bubonic plague, undergoes temperature-dependent hypoacylation activity in response to the increase of temperature to 37°C (Montminy et al., 2006). The hypoactylation allows for evasion of the host defences as the LPS Is unrecognisable (Montminy et al., 2006). Similarly, *Shigella* (the pathogen of focus for this thesis*)* species have evolved a remodelling mechanism for the LPS. During proliferation in epithelial cells *Shigella* drastically reduce the acylation rate of lipid A (Paciello et al., 2013). The evolution of this host adapted mechanism dampens immune surveillance and interferes with the processes involved in pathogen recognition. Furthermore, as mentioned within section 1.1.1.3.1, USA300, very successful strains of MRSA, were noted to have other elements contributing to its global success other than AMR. USA300 was noted to have an Arginine Catabolic Mobile Element (ACME) unique to USA300 (Vanhommerig et al., 2014). This ACME is thought to underlie the prolific biofilm formation capabilities of USA300, an important characteristic for transmission, survival and evasion of the innate immune system. contributing to its success (Vanhommerig et al., 2014). The ACME encodes for *speG* which allows USA300 strains to withstand levels of polyamines (e.g., spermidine) produced in skin that are toxic to other closely related *S. aureus* strains whilst speG-mediated polyamine tolerance also enhances biofilm formation, adherence to fibrinogen/fibronectin, and resistance to antibiotic and keratinocyte-mediated killing (Planet et al., 2013). The ACME clearly is a major contributor to the extraordinary

pathogenic success of the USA300 strains by enhancing virulence, immune evasion and survival (Vanhommerig et al., 2014).

The immune systems of eukaryotic cells have evolved into complex and multifaceted systems capable of combatting multiple pathogenic infections. In response however, historical pathogens have evolved to subvert the hosts immune defences and cause disease. Historical pathogens such as *Y. pestis* and *Shigella* have evolved highly effective and complex mechanism to evade, alter and interfere with immune defences contributing to their long-term success as pathogens.

1.1.1.3.3. Interbacterial competition: Having the edge

It is not only through human interaction that bacteria have evolved to resist or evade. Bacteria are frequently found in large and diverse communities in the environment and in the microbiome of a human host. These microbial communities are highly competitive with limited resources determining survival of the fittest. There has been clear evolutionary emergence of numerous competition mechanisms. The stabilisation of these mechanisms within bacterial genomes shows the true importance of these mechanisms in survival and infection.

Exploitative competition is defined as passive competition. Bacterial species deplete the surrounding nutrients and in doing so prevent competitors from utilising the resources effectively (Birch, 1957). Specialised metabolites (SMs) are molecules which are not involved in primary metabolism but are involved in other processes. These SMs were previously known as secondary metabolites due to their production in late stage growth in laboratory cultures, however, these metabolites may be essential for some bacteria to survive in

competitive environments (Price-Whelan et al., 2006). Multiple bacterial species have evolved metabolites for this purpose, for example the evolution of siderophores. Siderophores are metabolites produced by bacteria for scavenging iron from the environment (Hider and Kong, 2010). Iron is essential for multiple processes in bacteria especially for cytochromes and iron-sulphur proteins (Stubbendieck and Straight, 2016). There have been numerous examples of siderophore-mediated competition, as exemplified between the interbacterial competition of *S. aureus* and *Pseudomonas aeruginosa*. Both *S. aureus* and *P. aeruginosa* are human opportunistic pathogens typically associated with cystic fibrosis (CF) patients (Harrison et al., 2008). A study demonstrated that in the presence of *S. aureus,* siderophore production was upregulated in *P. aeruginosa* showing the competitive advantage that siderophores contribute to (Harrison et al., 2008).

In contrast to exploitative competition, interference competition is aggressive where bacteria release antagonistic factors produced to impede competitors (Stubbendieck and Straight, 2016). Bacteria have developed multiple strategies for this purpose, one such mechanism is protein secretion systems. The secretion systems act as a delivery mechanism for antibacterial toxins to target cells therefore killing or inhibiting growth of competitors (Lin et al., 2020). Among the nine secretion systems which have been identified so far there are five which have proven capability to deliver antagonistic compounds. One of the contact dependent systems, the type VI secretion system (T6SS) is a known mechanism found in numerous bacterial species for interbacterial competition (Schwarz et al., 2010). In the historical pathogen *V. cholerae* the T6SS' primary function is to aid interbacterial virulence. In one study it was shown that the T6SS was highly effective against multiple Gram-negative bacteria including *E. coli* and *Salmonella Typhimurium* (MacIntyre et al., 2010). These

bacteria are common within the human microbiome and so would be present as competitors during *V. cholerae* infection.

In particular for *Enterobacteriaceae* strains, the interbacterial virulence factor - colicins may play a key role. Colicins are small toxic proteins produced by many enteric bacteria and are usually encoded on small colicinogenic plasmids alongside the colicin lysis protein that is responsible for colicin release and the immunity protein that protects the host from its own colicins (Riley, 1993, Cascales et al., 2007). Typically, they are noted to be toxic against *E. coli*, a very well-known bacterial species in the human microbiome, and so are extremely beneficial during host infection when bacteria must outcompete their competitors in the microbiome.

The bacterial mechanisms which have evolved to aid interbacterial competition are numerous and diverse. They represent mechanisms of clear importance for the persistence and survival of bacteria to cause infection and have evolved to persist within pathogen genomes. These mechanisms clearly contribute to the success and persistence of historical pathogens.

1.1.2. Traditional methodologies

Now that I have given a general overview on how human interventions and immune systems have shaped bacteria, I will now focus on ways in which we as researchers can investigate long-term success of pathogens. To investigate what factors contribute to the long-term persistence of historical pathogens, a wide range of techniques and tools are required to conduct an in-depth investigation. Recent technological advancements have increased the availability and accessibility of biological data while also making investigations quicker and more comprehensive.

1.1.2.1. Whole genome sequencing

In 1977 a breakthrough in DNA sequencing technology altered the biological field forever. The development of Sanger's 'chain-termination' technique involved the selective incorporation of chain-terminating dideoxynucleotide by DNA polymerase during in-vitro DNA replication (Sanger et al., 1977). The Sanger method led to the first DNA sequence of the bacteriophage ΦX174 and quickly became the most common technology used to sequence DNA for many years (Sanger et al., 1977). Improvements to Sanger sequencing occurred in the following years. Most notably the replacement of phospho-radiolabelling with fluorometric based detection and improved detection through capillary-based electrophoresis allowed for the development of the first commercial DNA sequencing machines (Smith et al., 1985, Prober et al., 1987, Swerdlow and Gesteland, 1990). The first-generation of these DNA sequencing machines could only analyse a sequence read of less than one kilobase (kb) in length, with multiple runs required to read longer genomes (Anderson, 1981).

Today's complex genomic questions rely on a depth of information which would have been unattainable by traditional DNA sequencing technologies. In 2005 the first commercialised new next-generation sequencing (NGS) technology was released known as the pyrosequencing method by Roche (Margulies et al., 2005). The mechanics of NGS are similar to that of capillary electrophoresis with the exception of a few major improvements. Firstly, the replacement of bacterial cloning of DNA fragments with the preparation of NGS libraries in a cell free system. Secondly and possibly the most important improvement, is the sequencing and detection of fragments is completed in a massively parallel fashion, improving speed and accuracy whilst reducing overall cost (van Dijk et al., 2014). The advent of NGS immediately revolutionised the field of genomics. Illumina's NGS DNA technologies

are now considered the market leader of clinical research sequencing with highly accurate (over 99.9%) and inexpensive reads on a massive scale (Logsdon et al., 2020). However, there is a drawback to the NGS technologies speed and cost effectiveness, the relatively short read length (van Dijk et al., 2014).

Long-read sequencing or third generation sequencing can determine the nucleotide sequences of long sequence of DNA of up to between 10,000 and 100,000 base pairs, much longer than NGS technologies. The longer reads provide better accuracy for DNA containing repetitive regions, improve de novo assemblies and detect structural variants (Amarasinghe et al., 2020). Furthermore, long-read sequencing eliminates PCR amplification bias by sequencing native molecules of DNA and RNA allowing direct base detection of modifications such as methylation (Mantere et al., 2019). The continued development of long-read sequencing platforms has produced several technologies which are cost effective and high throughput for research purposes. Two long-read sequencing technologies have been elevated throughout the long read field, Pacific Biosciences' Single-Molecule Real-Time (SMRT) sequencing and Oxford Nanopores Technologies' (ONT) nanopore sequencing (Amarasinghe et al., 2020). Released in 2011 and 2014 respectively, these technologies have become a staple for whole genome sequencing. In 2015 there were 99 mammalian genomes, none of which were long-read sequenced, but in 2020, there have been more than 800 genome assemblies utilising either of these two technologies (Logsdon et al., 2020). These advancements are key to investigations into historical pathogens.

The recent improvements over the past two decades in sequencing technologies have enabled researchers to study microbial systems at a depth never seen before. NGS sequencing and long-read sequencing have become an established and increasingly used tool for the bacterial genomic research. With the continued development and ongoing cost

reduction there has already been use of sequencing technologies for more-routine applications such as for surveillance of global pathogens as seen with the SARS-CoV-2 pandemic (Balakrishnan, 2022).

1.1.2.2. Bioinformatic techniques

In tandem with the continuous improvements in sequencing technologies, the need for computerised tools and software to handle, manipulate and visualise sequencing data has led to a rise in the field of bioinformatics.

It is easy to imagine that modern bioinformatics was a recent development, coming to the aid of huge quantities of generated sequencing data. However, bioinformatics originated over 50 years ago, even before DNA could be sequenced (Gauthier et al., 2018). The original focus of bioinformatics grew from the publication of protein sequences, the first of which was insulin in 1953 (Sanger and Thompson, 1953). Margaret Dayhoff is commonly considered the 'mother of bioinformatics' and developed a computer program to predict primary protein structure from peptide sequencing data (Moody, 2004).

Today, bioinformatics is considered an interdisciplinary field combining biology, mathematics, information technology and statistics. With the exponential increase in biological data in public databases a plethora of tools have been developed to analyse them. There are now tools to analyse DNA, RNA and protein sequences from assembly to annotation to phylogenetic analyses and beyond (Mehmood et al., 2014). For assembly of microbial DNA sequences commonly used bioinformatic tools include Unicyler (Wick et al., 2017a), Velvet (Zerbino, 2010) and SPAdes (Bankevich et al., 2012). All of these tools have been resourcefully developed to overcome issues with tandem repeats by utilising paired end information making these tools invaluable for accurate assembly of sequencing data.

When there is combination of short and long reads, hybrid pipelines, such as Unicycler can be utilised to produce a complete genome (Wick et al., 2017a).

For functional genomics, the annotation step is key. Common tools include Prokka (Seemann, 2014) or the NCBI Prokaryotic Genome Annotation Pipeline (PGAP) (Tatusova et al., 2016). Accurate annotation is essential as sequencing of the genome only provides sequence information with no functional roles. Annotation works to provide structural features and functional roles of genes within the DNA sequence (Harbola et al., 2022). Annotated genomes can act as input to other bioinformatic tools, such as AMRFinderPlus (Feldgarden et al., 2019), which identify genes with the key functions of AMR and virulence. In 2022, there are now over 50 freely accessible bioinformatic resources for detection of AMR determinants including ResFinder, CARD and AMRFinder (Bortolaia et al., 2020, Alcock et al., 2020, Feldgarden et al., 2019). The accuracy of AMR phenotype prediction within well studied *Enterobacteriaceae* has been proven to be as high as 98.4% in a large dataset containing multiple *Enterobacteriaceae* (Feldgarden et al., 2019). These bioinformatic tools represent an invaluable resource for predicting virulence of key global pathogens.

A recent focus of bioinformatics has been to find ways to produce accurate in silico 3D structure of proteins. This has the advantage of predicting putative function without the need for crystallography which is a complex and time-consuming laboratory process. There are currently only approximately 144,000 unique protein structures which have been determined (2019b). However, this represents only a miniscule percentage of the estimated billions of known proteins. Protein structure solving is hindered by the complex and labour-intensive effort required to experimentally determine a single protein structure. The latest and cutting-edge bioinformatic tool is Alphafold, an AI system which predicts a protein's 3D structure from its amino acid sequence and frequently achieves accuracy competitive with

experimental values (Jumper et al., 2021). Alphafold addresses a 50-year protein folding problem where the tool can regularly provide protein structures with atomic accuracy even when there is no homologous structure present (Jumper et al., 2021). This computational approach represents an innovative step towards a rapid and convenient approach to understanding unsolved protein structures.

It is evident that there has been an exponential increase in the development of bioinformatic tools which has been stimulated by the vast amount of biological data available in public databases. It is essential to have tools that analyse, manipulate, and visualise this data to fully understand the structural and genetic components of bacterial genomes. In depth interrogations of bacterial genomes with these tools is aiding our understanding of why pathogens are so successful.

1.1.2.3. Genome wide association studies

In recent years there has been an escalation in the use of bioinformatic approach known as a bacterial Genome Wide Association Study (GWAS) which has been made possible by the development of new bacterial GWAS tools that account for the clonal population structure (Falush and Bowden, 2006).

A GWAS aims to identify variants at genomic loci (typically single-nucleotide polymorphisms (SNPs) or gene presence/absence) which are associated with complex traits in a population, such as an AMR phenotype or a disease phenotype (Visscher et al., 2012). The method surveys entire genomes for variants that occur more frequently in case isolates (people with a trait) than within the controls.

Since the first bacterial GWAS was published in 2013, there has been a vast increase in the number of published GWAS' resulting in over 6000 papers (both bacterial and human) being published in 2021 alone (according to PubMed)(Sheppard et al., 2013). As of September

2018, GWAS' were reported to have accelerated gene discovery, with a GWAS catalogue of over 5000 GWAS' and over 71,000 variant-trait associations (Buniello et al., 2018). This approach is becoming an established and valued tool in modern genetics. Today there are various tools which make utilising bacterial GWAS' accessible to all researchers such as Scoary, PLINK, SEER and Pyseer (Brynildsrud et al., 2016b, Purcell et al., 2007, Lees et al., 2016, Lees et al., 2018). Previously there were limitations in bacterial GWAS' due to the clonality of bacterial populations. The clonality forms a major problem especially when comparing lineages with different phenotypes, which leads to all variants that separate the lineages observed are seemingly associated with the trait (Lees and Bentley, 2016). Recent improvements in bacterial GWAS' now incorporate a correction for population structure helping to eliminate the clonality hurdle (Lees et al., 2018).

Bacterial GWAS' have already been demonstrated to have a profound impact on public health research. For example, in *Mycobacterium tuberculosis*, a GWAS by Farhat et al. (2019) aimed to identify variants related to the resistance to anti-TB drugs. Through analysis of over 1400 clinical *M. tuberculosis* isolates, 13 genetic loci associated with resistance to anti-TB drugs were identified (Farhat et al., 2019). Identification of these genetic loci, provides a better understanding behind genetic mechanisms of resistance for *M. tuberculosis* and could perhaps lead to more-target drug approaches in the future (Farhat et al., 2019). The example given above shows the importance of GWAS' and the positive impact it can have for management of public health pathogens.

Bacterial GWAS represents an imperative bioinformatic approach to better understand the biology of disease, under the hypothesis that better understanding of traits will lead to improved prevention, treatment, and management. When considering factors which have

contributed to the long-term success of a pathogen GWAS represents a vital approach to consider for various important traits.

1.1.3. Factors contributing to the long-term success of pathogens

Now that I introduced how I can investigate these long-term successful pathogens, I want to consider what factors I would expect to find from these methodologies. When considering why some pathogens are more successful in causing disease than others, several common factors need to be considered, such as AMR and virulence determinants. Each of these greatly contribute to a pathogens ability to survive, infect, and transmit disease in a population. However, bacterial genomes contain a plethora of genes whose function may be contributing to this success of a pathogen in ways which are yet to be explored. This represents an exciting opportunity to explore the 'black hole' of bacterial genomes.

1.1.3.1. Bacterial mechanism of antimicrobial resistance

As briefly discussed before, AMR is an essential factor for bacterial infection and transmission. In order to survive within human hosts many bacteria acquire multiple AMR determinants to persist and cause infection, such as *S. aureus* who through acquisition of several prominent AMR determinants has evolved to become a global 'superbug'.

The acquisition of AMR determinants into bacterial genomes is typically through horizontal gene transfer (HGT) either by transformation, conjugation or transduction of antimicrobial resistance genes (ARGs), efflux pumps and plasmids (Tenover, 2006). The exchange of genetic material among similar and diverse species of bacteria, particularly of MDR plasmids, has led to the emergence of 'superbugs'. In *Enterobacteriaceae* the dissemination of epidemic plasmids belonging to IncF with divergent replicon types FIA and FII is strongly associated with the pandemics caused by MDR *E. coli* and *K. pneumoniae* (Mathers et al., 2015, Carattoli, 2009). The pandemics are typically due to the spread of certain high risk

clones which contain the epidemic IncF plasmids, namely, *E. coli* sequence type 131 and *K. pneumoniae* ST258 (Mathers et al., 2015). The importance and interdependence of high-risk pandemic clones and acquisition of MDR plasmids has widely been reported as a key factor in global pathogen dissemination. For orientation of this thesis, the pKSR100 plasmid within *Shigella* has already been shown contribute to epidemiological success among *Shigella*, however, key *Shigella* plasmids will be discussed further later (Malaka De Silva et al., 2022). Susceptible bacteria may also acquire resistance through spontaneous mutations. These mutations can aid resistance in a variety of ways such as through upregulating the activity of efflux plumps, modifying the target protein or binding site that the AMR agent binds to or in some cases downregulating an outer membrane protein channel that is needed for AMR entry (Tenover, 2006). For example, quinolone resistance is most often the result of various chromosomal mutations in the bacterial gyrase and topoisomerase IV genes (Aldred et al., 2014). Quinolones typically act through converting gyrase and topoisomerase IV into toxic enzymes which detrimentally fragment the bacterial chromosome. Specific mutations within these gyrase and topoisomerase IV genes weaken the interactions between quinolones and the targets causing significant resistance (Aldred et al., 2014). Spontaneous mutations have also been observed that affect efflux pumps resulting in significant upregulation. Within *E. coli,* mutations in the repressor gene acrR lead to heightened expression of the acrAB MDR efflux pump resulting in an escalation of resistance (Piddock, 2006, Wang et al., 2017).

The acquisition and subsequent conservation of antibiotic resistance carries significant fitness costs. Typically these fitness costs are observed as reduced growth rate, invasiveness and interbacterial competition abilities (Vogwill and MacLean, 2015). Plasmids harbour multiple AMR genes which act through modification or efflux of antibiotics, mechanisms which may disturb bacterial growth (Dahlberg and Chao, 2003). Furthermore, plasmid

replication and gene expression is typically reported to have a negative impact on growth independent of the function of the plasmid (Dahlberg and Chao, 2003). Mutations conferring resistance also exhibit a fitness cost, partially due to their impact on function in key biological genes. For example, in *P. aeruginosa* resistance mutations in *nfxB* gene result in diminished type III secretion system-dependent (T3SS) virulence and impaired fitness in terms of motility (Stickland et al., 2010).

Regardless of the fitness costs associated with the acquisition and subsequent conservation of AMR determinants, bacteria frequently show continued evolution towards greater resistance and more resistance mechanisms. These AMR determinants are clearly essential for survival, infection and transmission in an antibiotic treatment driven modern society making them non-negotiable genes within bacterial genomes. It is clear that AMR determinants would be observed as key factors associated with the long-term success of pathogens.

1.1.3.2. Virulence factors

In a similar fashion to AMR, the continued evolution and acquisition of virulence factors has been observed in bacterial genomes. The evolution of virulence within pathogens presents some ambiguity. Pathogenic species are often solely dependent on their host species for successful transmission, however, frequently the host must be damaged during this dependence. A broad spectrum of virulence mechanisms has nonetheless been positively selected for due to the wide array of other pathogenic species available for HGT coupled with host immune defences.

P. aeruginosa is a prime example of a pathogen where a broad virulence set has greatly bolstered its capability to be the dominant pathogen within CF patients. *P. aeruginosa* has a

large bacterial genome (5 – 7 MB) that contributes greatly to its adaptive capacity and metabolic flexibility (Jurado-Martín et al., 2021). Depending upon the environmental stressors its virulence repertoire (Figure 1) allows the pathogen to alter its state resulting in successful acute and chronic infections (Jurado-Martín et al., 2021). Within the extensive repertoire there are components acquired to disable the host immune system, most notably the type III secretion system (T3SS), to aid attachment to human epithelial cells, adhesins and biofilm formation genes, and to aid interbacterial competition, three quorum sensing systems (Figure 1) (Jurado-Martín et al., 2021). The acquisition of the number and broad capabilities of the virulence repertoire directly contributes to *P. aeruginosa* as the successful and dominant pathogen in CF patients.

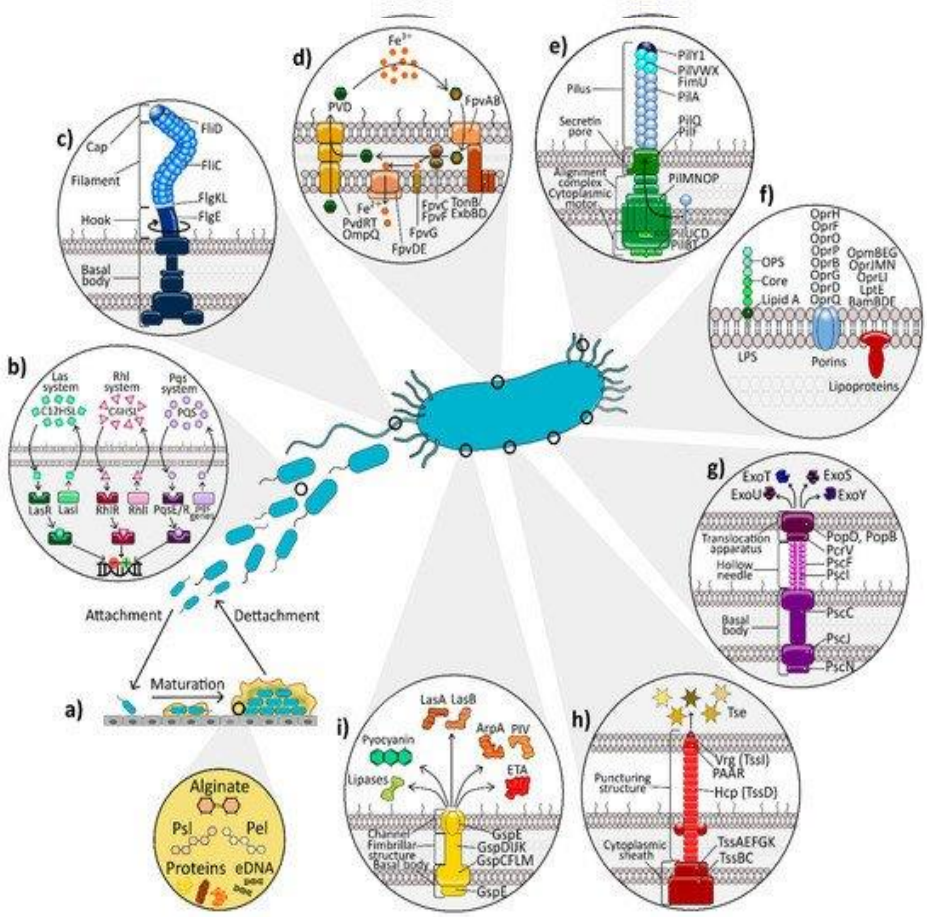


Figure 1: The wide repertoire of virulence factors utilised by *P. aeruginosa*.

a) Biofilm formation, b) Three main quorum sensing systems, c) flagellins, d) Siderophore uptake system, e) Type IIII pili, f) LPS and outer membrane proteins, g) Type III secretion system, h) Type VI secretion system and i) Type II secretion system **(Figure from Jurado-Martín et al. (2021))**

Within other Gram-negative bacterial species, virulence factors have also been demonstrated as fundamental to successful infection. *Shigella*, the causative agent of bacillary dysentery, has a plethora of virulence factors crucial to successful host infection. Firstly, *Shigella* secretes factors which induce the classic watery diarrhoea seen in early infection (Mattock and Blocker, 2017). The *Osp* genes, encoding T3SS effector proteins, modify mitogen-activated protein kinase (MAPK) pathways contributing to initial inflammation (Zurawski et al., 2009). In conjunction with Osp proteins, *Shigella* enterotoxins 1 and 2 mediate early fluid secretion in the jejunum to establish infection in the colon and produce the characteristic watery diarrhoea seen early in shigellosis (Mattock and Blocker,

2017, Vargas et al., 1999). Secondly, *Shigella* utilises a T3SS, located on the virulence plasmid, to inject virulence factors directly into the epithelial cells. The crucial contribution of the T3SS results in successful invasion of the epithelial cells providing a replicative niche (Muthuramalingam et al., 2021). In the case of *S. dysenteriae*, secretion of the shiga toxin may cause severe bloody diarrhoea (Mattock and Blocker, 2017). Lastly, *Shigella* continues to produce multiple virulence pathogens to down-regulate infection and evade host immune defences. For example, secretion of T3SS effectors OspG/I/Z dampens the host inflammatory response through inhibition of NFκB activation pathways (Sanada et al., 2012, Newton et al., 2010). The broad range of virulence factors acquired by *Shigella* species are imperative to all stages of *Shigella* infection, replication, and transmission.

The above examples give a general overview of some of the common mechanisms found within bacterial species which are key to virulence and pathogenic success. Due to this thesis' focus on the *Shigella* species, more in depth explanation of key *Shigella* virulence factors will be discussed within 1.3, 1.4 and 1.5.

Many bacterial species have acquired a broad range of virulence factors which are essential for various stages of successful invasion, replication, and transmission. These factors aid evasion of the host immune system, invasion of host cells and outcompeting other bacteria within their bacterial niche. Without these key genes pathogens could not persist and cause successful disease globally for prolonged periods of time.

1.1.3.3. Other factors: The unknown or under-researched

The improvements in DNA sequencing have produced a huge collection of biological data which represents an untapped resource to investigate what makes pathogens so successful over time. There are factors which researchers already know are imperative to successful

infection such as AMR and virulence factors. However, apart from these frequent, accepted factors there are a plethora of genes with unknown function which could be contributing to the long-term persistence of bacterial species.

Genes with other key biological functions such as aiding nutrient mining or phage related genes could be contributing to the success of pathogens. A prophage is a bacteriophage genome that is integrated into the circular bacterial chromosome or exists as an extrachromosomal plasmid within the bacterial cell. In Shiga toxin-producing *E. coli* (STEC), the *stx* genes (a prominent bacterial toxin) are located in the sequence region of stx prophages (Zhang et al., 2020). This is an example of prophages having key functions in some pathogens virulence capabilities, but many remain uncharacterised and unannotated.

In addition to lesser-known factors, there is a large proportion of genes within bacterial genomes which are known as 'hypotheticals'. These are uncharacterised genes which are not currently annotated and are of unknown function. Unannotated genes in global pathogens still play important roles in successful infection phenotypes. Within a reference strain of *E. coli* K-12 it has been estimated that approximately 35% of the genes lack confirmed function and some lack any detail whatsoever (Ghatak et al., 2019).

These uncharacterised factors whether they be lesser studied or completely hypothetical represent a unique and truly exciting opportunity to explore novel factors behind the long-term success of global pathogens. Being able to elucidate what contributes to the persistence of pathogens would have far reaching implications for public health through prevention and management.

1.2. A historical perspective

From section 1.1, I hope it is now evident how important these long-term persistent pathogens are. In order to fully understand how these historical pathogens are still so

successful a full evolutionary arc must be understood – from historical to modern day. Incorporating historical isolates into comparative genomic analyses facilitates a unique perspective. Historical isolates span key periods of time such as the pre-antibiotic era elucidating unique insights into a key era of history that can be extrapolated to understand how implementation of antibiotics has changed bacterial genomes. Tracking bacterial genomes over time could aid prevention, treatment, and management of key public health pathogens.

1.2.1. Historical isolates

The advancements in sequencing technologies have been phenomenal with technology advancing to a scale where the United Kingdom Health Security Agency (UKHSA) now implements routine whole-genome sequencing for surveillance (Grant et al., 2018). In this context, modern bacterial collections are utilised for rapid outbreak detection and intervention targeting. However, this technology can be utilised for much earlier isolates in time, and by analysing historical bacterial genomes researchers can gain unique insight into some of the world's global pathogens.

Historical bacterial genetic material is available from a plethora of different source types, from bones and teeth to historical isolate collections held in pathogen repositories. Teeth and bones provide a haven for bacterial DNA to survive and so provide a unique source of genetic material. This historical DNA is known as ancient DNA (aDNA). In one publication by Drancourt et al. (1998) DNA extracted from dental pulp was identified as *Y. pestis*, the causative agent of the plague. The extracted aDNA confirmed the presence of the plague in France during the 16th Century and was the first study highlighting dental pulp as a suitable media for confirming ancient septicaemia (Drancourt et al., 1998). Despite successes, there are challenges associated with utilising aDNA including degradation and contamination

(Knapp and Hofreiter, 2010, Gorge et al., 2016). However, these challenges are greatly aided by advancements in sampling strategies and the improvements in short-read sequencing allowing use of low concentrations of aDNA. aDNA represents a unique and exciting resource for retrospectively diagnosing historical diseases, thus aiding the understanding of transmission and dissemination of historical pathogens. In tandem with aDNA, there are also individual historical isolates isolated from later centuries which have less limitations than aDNA but still provide historical context.

Taking into the account the insight to be gained from individual historical isolates, it would be sensible to envisage that with a greater quantity of isolates an even more thorough historical picture could be discovered. Although many of these pathogen repositories are limited to the past few centuries, they still span many major historical and geographic events most notably wars and the implementation of antibiotics. There are several large global pathogen repositories. One such repository is the National Collection of Type Cultures (NCTC) held by UKHSA. The NCTC holds over 5000 type and reference strains, many of which are of medical and scientific research importance. Other prominent reference collections are Collection de l'Institut Pasteur (CIP), started in 1892 and currently housing over 12,000 bacterial strains, and the American Type Culture Collection (ATCC) started in 1925. These repositories represent a wealth of historical biological data to investigate the evolution of bacterial genomes.

Advancements in sampling strategies and NGS have heightened access to historical bacterial DNA which was previously unattainable. Through extraction from bones and teeth to curation of various historical isolate collections, there is a now a wealth of strains which together offer a unique insight into key time periods for bacterial evolution.

1.2.2. *The utility of sequencing historical isolates*

Through recent advancements in sequencing historical isolates have become much more accessible. Various types of historical isolates have already been used to gain insight into many fields of biology including epidemiology, bacterial evolution, and genetic diversity (Figure 2). Through complex analysis and combination of historical and modern data collections researchers can begin to disentangle the complex histories of global public health pathogens.

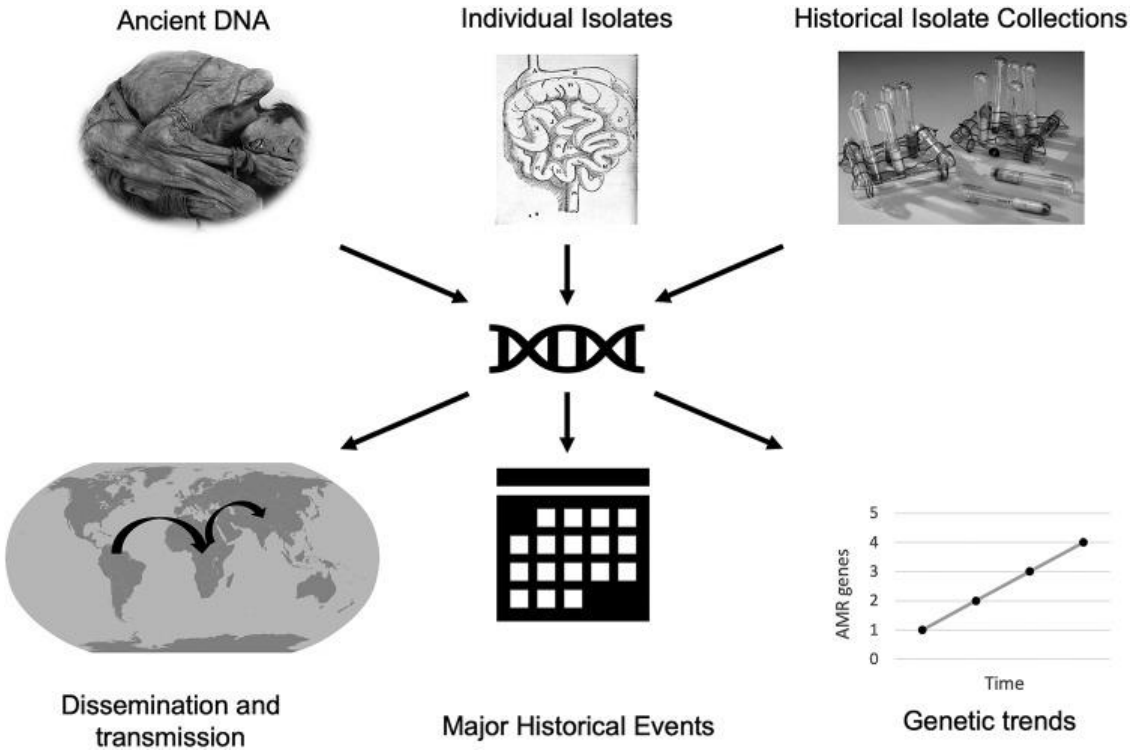


Figure 2: The utility of historical isolates

Schematic overview of the three types of historical isolates, illustrating some of the insights that can be gained through sequencing of historical isolates **(Figure from Bennett and Baker (2019)).**

One of the most innovative uses of historical isolates is to conduct analyses where historical isolates are used in combination with the plethora of modern isolate collections which are available in the modern day. An exquisite example of this was undertaken by Baker et al. (2014) where whole genome sequencing of an extant historical isolate of bacillary

dysentery, NCTC1, was undertaken. NCTC1 was isolated in 1915 from a British Forces soldier and represents the oldest extant isolate of *Shigella flexneri* in existence (Baker et al., 2014). Comparative genomics of NCTC1 with reference genomes of *S. flexneri*, *S. sonnei*, *S. boydii*, *S. dysenteriae* and *E. coli* revealed NCTC1 to be from the 2a sublineage of *S. flexneri*, a still prevalent lineage in modern day (Baker et al., 2014). Acquisition of genomic islands relating to AMR, virulence, and immune evasion were observed highlighting the areas which are key to bacterial evolution, alongside an island of unknown function. This study represented a prime example of the insight historical isolates can give through genetic comparison to contemporary isolates. Sequencing of the extant *S. flexneri* isolates provided a benchmark for the study of sublineage 2a, a still prevalent *S. flexneri* lineage.

Many historical isolate and collections span key time periods, including the discovery and implementation of antibiotics, making them immensely useful for demonstrating the evolutionary dynamics of bacteria especially for genes which contribute to the long-term success of pathogens such as AMR. With the continued transmission of many public health pathogens coupled with increasing AMR, it is important now more than ever to analyse bacterial evolution of these traits to better advise on future treatment and management of these key pathogens.

1.2.3. The Murray Collection

As previously discussed, historical collections are a key and unique resource for bacterial genetic insight due to their idiosyncratic trait of spanning large time periods. One such collection of particular interest for public health is the Murray Collection. The Murray Collection was amassed by the late esteemed microbiologist Professor Everitt George Dunne Murray and consists of over 600 bacterial strains (JB, 1965). The collection comprises of historical *Enterobacteriaceae* isolates spanning the pre-antibiotic era (1917 to 1954) from

diverse geographical locations (Baker et al., 2015a). Originally these isolates were stored on agar slopes, however, in the 1980s REG Murray transferred the cultures to the NCTC at UKHSA, where they are still held today. The Murray Collection's 683 strains were primarily from *Salmonella, Escherichia, Shigella, Klebsiella* and *Proteus* genera and comprised a large diversity of subspecies and serotypes (Baker et al., 2015a). To confirm the genetic diversity and robustness of this profound collection 370 strains were whole genome sequenced and these genomes are now publicly available to provide a unique and exciting resource spanning the pre-antibiotic era (Baker et al., 2015a). Phylogenetic analyses and genotypic AMR profiles were determined for the collection curating a significant gnomic resource for the study of evolution and emergence of AMR in *Enterobacteriaceae*.

Valuable insights from the Murray Collection have already been gained. Seminal studies into conjugative plasmids were investigated utilising the Murray Collection (Datta and Hughes, 1983, Hughes and Datta, 1983). In past decades, conjugative plasmids encoding AMR determinants had become common in *Enterobacteriaceae*. Through comparison of the Murray Collection 'pre-antibiotic' plasmids and contemporary isolates, it was observed that most 'pre-antibiotic' plasmids belonged to the same incompatibility groups as modern resistance plasmids (Datta and Hughes, 1983). The use of historical isolates showed that modern resistance is acquired through insertion of new genes on existing plasmids, a profound insight into the emergence of AMR. Furthermore, a similar finding was observed for mercury resistance within *Salmonella* based upon the historical insight of the Murray Collection *Salmonella* isolates (Essa et al., 2003, Jones and Stanley, 1992).

In addition to these seminal studies, a large study by Wand et al. (2015b) involving the Murray Collection's *K. pneumoniae* isolates aimed to characterise the pre-antibiotic isolates in terms of antibiotic/disinfectant susceptibility combined with virulence in a model

organism. Susceptibility testing revealed over 30% of historical isolates were already resistant to penicillin, through the bla_{SHV} β-lactamase gene, prior to the introduction of penicillin use clinically (Wand et al., 2015b). For both disinfectant resistance and virulence, the historical isolates exhibited lower rates of disinfectant resistance and virulence when compared to modern *K. pneumoniae* isolates (Wand et al., 2015b). The different array of virulence factors identified in the historical isolates compared to the modern isolates suggests acquisition of other key virulence factors is necessary for more virulent phenotypes such as biofilm formation which was observed in 5 times more modern isolates than historical isolates (Wand et al., 2015b). The characterisation of the historical *K. pneumoniae* provides significant insights into the adaptation and evolution of this public health pathogen into a more virulent and resistant pathogen.

The Murray Collection represents a large collection of diverse bacteria spanning the 'pre-antibiotic' era, now publicly available from the NCTC. This unique and invaluable collection is a profound resource for the study of bacterial evolution and emergence of key traits. The studies which have previously utilised the Murray Collection have demonstrated the valuable insights to be gained from using a historical isolate collection combined with the plethora of modern isolates available. These studies represent only the 'tip of the iceberg' for the potential of the Murray Collection and many of the evolutional insights remain entirely untapped for further scientific research.

1.3. *Shigella*: A prominent public health pathogen

Shigella species are the primary focus of this thesis and so therefore this public health pathogen will now be discussed in further detail.

1.3.1. Shigella*: The bacteria*

Shigella was first discovered in 1896 by Kiyoshi Shiga, when investigating an outbreak of bacillary dysentery in Japan. Bacteria of the genus *Shigella* have a lengthy and prolonged history as a public health pathogen, clinically manifesting itself as shigellosis. Shigellosis is typically characterised by fever, nausea, abdominal cramps, and bloody diarrhoea. Diarrhoeal disease is still a major global health issue and today accounts for approximately 1.3 million deaths each year (Estimates, 2016). Of greatest concern is diarrhoeal disease occurring in young children where over 950 million cases of diarrhoea occur annually, especially in LMIC (Estimates, 2016). *Shigella* species represent the second leading cause of diarrhoeal disease globally and recent estimates attribute 125 million cases and 216,000 deaths annually to shigellosis (Troeger et al., 2018, Khalil et al., 2018). 99% of *Shigella* infections occur in LMIC with 69% of infections in the paediatric population (Williams and Berkley, 2018). Coupled with the highly infectious nature of *Shigella* species, the WHO has recognised *Shigella* as a priority organism in terms of AMR due to its ever-expanding AMR determinant repertoire (Shrivastava et al., 2018). Today, *Shigella* remains a highly infectious disease-causing global infections and so endures as a prominent public health pathogen.

1.3.1.1. Biology

The genus *Shigella* are Gram-negative, facultative anaerobes, commonly known as the primary agent of bacillary dysentery. These Gram-negative bacilli are short, only 0.5 p x 1-3 p in size, which lack motility, have no spore-forming abilities, and typically are non-capsulated, however, *S. sonnei* has been recently shown to have a group IV capsule (Hale and Keusch, 1996, Caboni et al., 2015). *Shigella* spp are intracellular pathogens, defined as pathogens capable of evading the immune system and persisting and replicating within host cells. An optimum temperature of 37°C is well suited to the human host and can survive in harsh environments such as freezing temperatures and low pH (4.5) for several days

(Bhunia, 2018). *Shigella* are differentiated form the closely related *E. coli* based on pathogenicity (acquisition of the virulence plasmid), physiology (failure to ferment lactose), and serology.

1.3.1.2. Shigellosis

The genus *Shigella* is known as the causative agent of shigellosis, a historical disease which is still prevalent in modern society. Although infections occur globally and people of all ages may become infected, the majority of the disease burden occurs among children under the age of 5 living in LMIC (Kotloff et al., 2018). Current estimates reveal *Shigella* spp role as a prevalent modern-day pathogen with over 200,000 deaths estimated every year (Khalil et al., 2018).

The primary treatment for shigellosis relies on rehydration and anti-diarrhoeal medications with antibiotics being prescribed when necessary. Current WHO guidelines support the use of fluoroquinolones as the first line treatment for shigellosis, followed by β-lactams and cephalosporins, and/or azithromycin (Williams and Berkley, 2018). For the treatment of MDR strains of *Shigella*, pivmecillinam and ceftriaxone are reserved as a final treatment option but are limited due to their expense and formulation (Williams and Berkley, 2018). Current guidelines have not changed extensively in response to the increasing AMR observed within some *Shigella* strains. The future of successful treatments of shigellosis may be majorly hindered by the continued AMR acquisition into bacterial genomes and so close monitoring is needed.

1.3.1.2.1. Source and transmission

The primary source of shigellae is thought to be through human to human transmission through the faecal-oral route, although there are cases of faecal contaminated drinking water and food (Lamps, 2009). Shigellosis is a highly infectious disease with an extremely

low infectious dose (10-100 cells) required for successful infection. In LMIC this poses a troubling dilemma due to the lack of adequate sanitary facilities resulting in elevated case rates (Bhunia, 2018). Children under the age of 5 typically show higher case rates due to their inferior hygiene practices, however, adults are still susceptible even with their heightened immunity and awareness and access to hygiene.

1.3.1.2.2. Pathogenesis and invasion

In contrast to many other invasive pathogens that reside in host cells, *Shigella* species not only invade the colonic mucosa but also replicate and disseminate within the mucosa instead of penetrating deeper into the tissue for survival (Carayol and Tran Van Nhieu, 2013). *Shigella* species are able to invade the colonic and rectal epithelium (the clinical characteristic of shigellosis) causing infection of the colonic mucosa where severe tissue damage occurs leading to abscesses and ulceration (Jennison and Verma, 2004). This damage and destruction of the epithelial layer is what causes the symptoms of bacillary dysentery – bloody diarrhoea and severe abdominal pain (Jennison and Verma, 2004). *Shigella* spp once ingested show resistance to low pH (pH2.5 for over 2 hours) resulting in survival during transit through the stomach. One of the early essential steps in successful infections is adhesion which must first take place before all other intracellular steps occur. There are multiple traditional adherence factors including *csg* and *fim* genes as well potentially unannotated novel genes which have yet to be characterised as there is still some mystery into the exact adherence mechanisms (Chanin et al., 2019a). However, it is true that without adhesion *Shigella* species would not be successful intracellular pathogens. Invasion of the colonic epithelium is thought to be associated with an intense inflammatory host response (Figure 4). On inflammation, macrophages are attracted to the bacterial site and ingest *Shigella*. *Shigella*-induced pyroptotic death of macrophage leads to the

recruitment of polymorphonuclear cells and contributes to the destabilization of the epithelial layer (Carayol and Tran Van Nhieu, 2013). This destabilisation results in successful *Shigella* invasion at the levels of the Peyer's patch, through M cells (Figure 3) (Sansonetti and Phalipon, 1999). Following successful invasion, *Shigella* are able to lyse the phagocytic vacuole, and this permits intracellular replication. Dissemination to adjacent cells is achieved via polymerizing actin at one bacterial pole forming comet tails which protrude and invade into adjacent cells. Infected cells with high bacterial load from replication underdo *Shigella*-induced death via proinflammatory necrotic death and membrane damage (Carneiro et al., 2009).

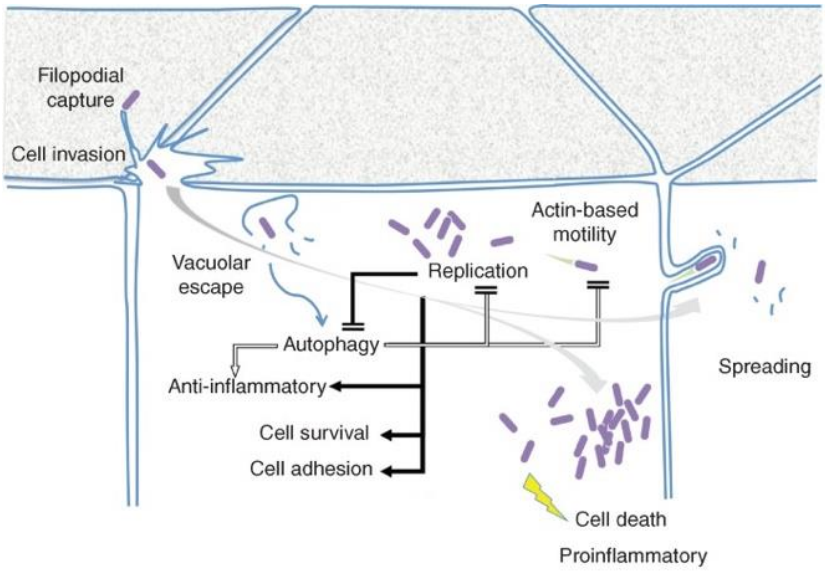


Figure 3: *Shigella* Invasion

Graphical overview of shigella invasion, intracellular replication, and spreading in epithelial cells **(Figure from Carayol and Tran Van Nhieu (2013)).**

A wide range of virulence factors are vital for the *Shigella's* invasion of the colonic epithelium. Most notably the virulence factors found on the 30kb pathogenicity island on the virulence plasmid (pINV) (Pilla et al., 2017). This island encodes for a T3SS which is vital for delivery of effector proteins such as IpaA, IpaC, IpgD and VirA (Bhunia, 2018). These effectors are involved in the induction of membrane ruffling to mediate micropinocytosis to

allow successful bacterial entry to host cells. In addition to the pINV, located on the chromosome are key factors involved with the plasmid-encoded virulence genes such as *virR* gene controlling temperature-dependent expression of the T3SS Spa/Mxi proteins (Bhunia, 2018). Further to the genes required for successful entry Shigellae produce three types of enterotoxins depending on the *Shigella* strain (Faherty et al., 2012). Each enterotoxin is secreted in the intestine and contributes to the watery diarrhoea symptoms observed through during shigellosis. In rabbit models two *Shigella* enterotoxins have been shown to increase water accumulation in the ileal loop model, supporting their known function as a contributor to diarrhoea (Faherty et al., 2012)

1.3.1.3. Classification

The genus *Shigella* is split into 4 species: *S. flexneri*, *S. sonnei*, *S. dystenteriae* and *S. boydii*. These four species are then subdivided into serotypes based on variation in their O-antigen and biochemical characteristics (Sahl et al., 2015). *S. dysenteriae* divided into 15 serotypes, *S. flexneri* and *S. boydii* into over 20 serotypes each and *S. sonnei* strains belonging to a single serotype (Hale and Keusch, 1996). There are three species of this genus which are the major disease-causing species: *S. flexneri, S. sonnei* and *S. dysenteriae*. *S. flexneri* and *S. sonnei* account for most shigellosis infections and will be discussed in detail in subsequent sections as they are the primary focus of this thesis.

There have been multiple recorded dysentery epidemics associated with warfare, for example during Napoleon's retreat from Moscow and the attacks on the Gallipolli peninsula during WW1 (Manson-Bahr, 1942). These epidemics were caused by the bacterium *Shigella dysenteriae* type 1 (Sd1), a bacteria which causes severe disease through secreting the cytotoxic shiga toxin (Manson-Bahr, 1942). Within the second half of the 20th century large outbreaks of Sd1 still occurred with death tolls reaching 20,000 in a 1969-1973 outbreak in

Central America (Njamkepo et al., 2016a). Within developing countries, *S. dysenteriae* is a major concern and the manifestation of the Sd1 infections is more severe because of its capacity to produce Shiga toxin (Stx) (exclusively produced by this type) (Talukder et al., 2003). The Stx is one of the most potent bacterial toxins known to mankind and is found within *S. dysenteriae* 1 and in some serogroups of *E. coli* (Melton-Celsa, 2014). Since the 1960s, *S. dysenteriae* has been associated with Latin America, Africa, India, and Bangladesh, causing outbreaks with high rates of morbidity and mortality. Although Sd1 infections are associated with severe symptoms and higher mortality rates, Sd1 infections are not the most prominent species and so quantitively are responsible for less cases.

The *Shigella* species *S. boydii* can be divided into 20 serovars with the 20th serovar being a relatively new and an emerging one (Kalluri et al., 2004). *S. boydii* infections are uncommon with this species only accounting for 2-6% of shigellosis globally (Nygren et al., 2013). *S. boydii* is therefore not a global public health concern. Phylogenomic analysis utilising Global Enteric Multicenter Study (GEMS) and other *S. boydii* genomes split the *S. boydii* isolates into three clades with clade 1 potentially diverged from clades 2 and 3 at an earlier point in the evolution of *S. boydii* (Kania et al., 2016). A core genome of 2230 genes was identified which were genetically similar in all isolates but each clade had a set of unique genes such as zinc-binding proteins for clade 1 or some phage components for clade 2 (Kania et al., 2016).

All of the four *Shigella* species contribute to the global burden of shigellosis. For the purposes of this project the species, *S. dysenteriae* and *S. boydii* have only been described briefly due to the focus of this thesis being solely on *S. sonnei* and *S. flexneri* which will be described in detail in future sections. *Shigella* species remain of public health concern and a prominent public health pathogen.

1.3.2. Genomic characterisation

The advancements in sequencing and subsequent manufacture of genomics tools have been utilised thoroughly in the investigation of *Shigella*. The technological revolution has revealed the dynamic genome flexibility of *Shigella* spp as well as shigellae's close evolutionary relationship with *E. coli* (The et al., 2016). Typically, *Shigella* genomes are over 4Mbp in length and comprised of a chromosome with various plasmids associated with each species and serotype. Although there are key genomic differences between the four species, there are some common genomic features.

1.3.2.1. The virulence plasmid

The acquisition of the *Shigella* pINV represents a key event in the formation and differentiation of the four *Shigella* spp from *E. coli*. The pINV is a large plasmid, 230kb, and encodes elements essential for virulence (Figure 4). pINV is a single copy, non-conjugative element which comprises of virulence-associated genes and plasmid maintenance genes separated by insertion sequence elements (IS) (Pilla et al., 2017, Sansonetti et al., 1982). Most virulence associated genes are located on a 30kb region known as a pathogenicity island (PAI). Encoded within the PAI include genes for epithelial entry and macrophage apoptosis as well as for a T3SS, essential for the invasiveness of *Shigella* (Bhunia, 2018). Along with the plethora of virulence factors there are typically systems to prevent plasmid loss. Within *S. flexneri* there are two partitioning systems, ParAB and StbAB, to ensure each daughter cell retains a copy of the plasmid as well as three functional toxin:antitoxin systems (TA systems) (Pilla et al., 2017). Typically, on the pINV TA systems are type II which consists of a toxic protein and an antitoxin antidote. In the presence of the plasmid, the antitoxin is expressed counteracting the activity of the toxin. However, in cells where the plasmid is lost the unstable antitoxin is degraded and is unable to interfere with the toxin

resulting in post-segregational killing (Pilla et al., 2017). Specifically the MvpAT TA system is essential for *Shigella* infection (McVicker and Tang, 2016, Dienemann et al., 2011). MvpAT confers plasmid maintenance at 37°C, the temperature of *Shigella's* invasion of the human host, and so allows successful T3SS injection of effector proteins (McVicker and Tang, 2016). Together all these elements encoded on the pINV confers the essential quality of this mobile genetic element for *Shigella* infection.

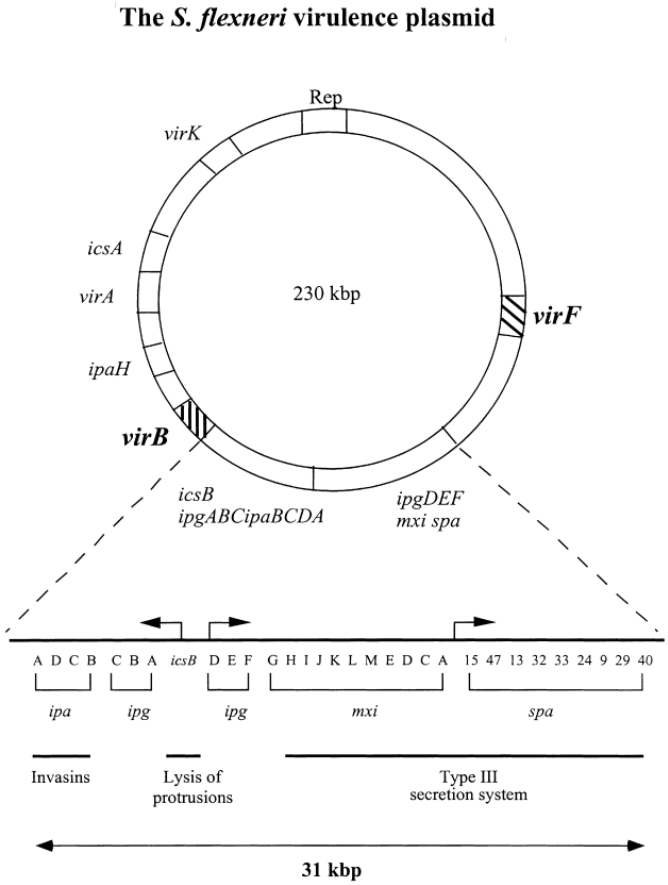


Figure 4: The *S. flexneri* virulence plasmid (pINV)

Representation of the *S. flexneri* virulence 230 kbp Virulence plasmid showing important genes including the T3SS and Ipa proteins (Figure from Dallman et al. (2016)).

Possibly the most important feature encoded on the pINV is the T3SS (Figure 4). Many Gram-negative bacteria encode T3SS to initiate contact with eukaryotic cells and manipulate the host cells to benefit the bacterium (Muthuramalingam et al., 2021). The T3SS is essential for virulence within *Shigella* and could be considered the most important virulence feature.

The T3SS is an injectosome composed of about 20 proteins that when assembled spans both bacterial membranes and the cell wall and is topped with a needle (Bajunaid et al., 2020). The T3SS apparatus is a complex nanomachine composed of 3 segments: a transmembrane basal body, the extracellular needle and the tip complex and the cytosolic complex (Bajunaid et al., 2020). Together these structures work to enable injection of bacterial effector proteins directly into the cell cytoplasm (Coburn et al., 2007). These effector proteins of the T3SS are involved in various virulence roles including interference with the host cell cytoskeleton to promote attachment and invasion, interference with cellular trafficking processes, cytotoxicity and barrier dysfunction, and immune system subversion (Coburn et al., 2007, Bajunaid et al., 2020). With these essential functions it is obvious why the T3SS helps to make pINV a non-negotiable MGE for the success of *Shigella* species.

There are many other plasmids found within *Shigella* species, possibly most notably is pKSR100. The selective pressure of antibiotic usage in treatment of sexually transmitted infections (STIs) such as gonorrhoea, common co-infections among shigellosis affected MSM, had led to the emergence of the globally successful sublineage of *S. flexneri* 3a following the acquisition of pKSR100 (Baker et al., 2015c, Malaka De Silva et al., 2022). pKSR100 carries azithromycin resistance through *mphA* and *ermB* genes, enhancing the epidemics and contributing to success. This is another key example of the importance of plasmids to the success of *Shigella* species.

Plasmids such as pINV and pKSR100 demonstrate the essential nature of plasmids to the virulence and success of *Shigella* species. Bacterial pathogens rely on the acquisition of plasmids to adapt to changing environments carrying advantageous traits necessary for survival, invasion and virulence. These MGE deserve proper consideration for their contribution to bacterial success, especially within *Shigella* species.

1.3.2.2. Insertion sequence elements

IS elements are highly abundant within all *Shigella* spp. And are defined as small transposable DNA sequences which insert themselves at many different sites within genomes (Siguier et al., 2014). It has been observed that many pathogenic genomes have undergone IS expansion accompanied by gene inactivation, genome rearrangement and reduction. IS expansion is commonly observed in bacteria that have adopted host-restricted lifestyles such as *Shigella* spp (Siguier et al., 2014).

Shigella spp are genetically similar to *E. coli* but converged to become human-adapted intracellular pathogens (Hawkey et al., 2020). IS elements are highly prevalent in all species of *Shigella* causing significant genetic variation in *Shigella* genomes. Within a reference genome of *S. flexneri* it was observed that IS elements were biased to non-coding regions of the genome, 28% higher than in other regions (Zaghloul et al., 2007). Furthermore, selection against genes encoding functions for signal transduction and cell motility were disrupted by ISs (Zaghloul et al., 2007). Hawkey et al. (2020) investigated the impact of IS elements on convergent evolution in *Shigella* species, *S. dysenteriae Sd1*, *S. sonnei* and *S. flexneri* and found that five IS elements (IS1, IS2, IS4, IS600 and IS911), in particular, have undergone IS expansion within each population resulting in convergent patterns of functional gene loss and genome degradation (Hawkey et al., 2020). IS expansion and corresponding genome degradation is most advanced in *S. dysenteriae*, however, convergent loss of *E. coli* metabolic capabilities and genome streamlining is being observed in a similar trajectory in both *S. sonnei* and *S. flexneri* (Hawkey et al., 2020). The importance of IS elements is clear for the genome streamlining and host adaption of *Shigella* spp.

1.3.2.3. Serotype switching

Shigella spp are further split into serotypes based on the LPS O-antigen structure. However, shigellae have evolved to acquire outside genetic material through various HGT mechanisms and this can lead to the rapid emergence of serotype variants through O-antigenic switching (Das and Mandal, 2019). This phenomenon is known as serotype switching and has been poorly quantified for many pathogens including *Shigella* (Ye et al., 2010, Allison and Verma, 2000). For *Shigella* spp this is of particular concern for vaccine development as the diversity of serotypes already complicates development but coupled with immune escape via serotype switching vaccine development becomes a substantial risk.

Bengtsson et al. (2022) performed an in-depth genomics analysis of over 1200 strains sampled as part of the GEMS, a modern isolate collection from between 2007 and 2011. During these analyses the phenomenon of serotype switching was investigated for *S. flexneri.* To qualitatively determine serotype switching, switching events were defined as when a new serotype emerged that was distinct from the majority serotype within a specific genotype (Bengtsson et al., 2022). Although switching events were revealed to be relatively infrequent (only 3.3% of isolates), the switching was facilitated by mutations, and gene presence/absence of various phage-encoded genes such as *gtr* and *oac* (Bengtsson et al., 2022). Even with the infrequent switching events the potential plethora of variants which could contribute to serotype switching poses complications to vaccine development. Furthermore, an estimated mean timeframe for switching events was observed at approximately 348 days means that switching events could impact the long-term effectiveness of vaccines especially when coupled with the selection pressure vaccines would provide (Bengtsson et al., 2022).

1.4. An introduction to *S. flexneri*

Shigella species have been estimated to cause over 165 million new infections globally per year, the vast majority of which occur in LMIC in children under the age of 5 (Ram et al., 2008). *S. flexneri* represents the most prevalent species of *Shigella* and is endemic to many LMIC and quantitatively is responsible for the majority of *Shigella* infections globally (Von Seidlein et al., 2006). Due to its geographical focus in LMIC, and additional virulence determinants including enterotoxins SHI-1 and SHI-2, *S. flexneri* infections are associated with higher rates of fatalities (Ud-Din et al., 2013).

1.4.1. Population structure

Prior to 2015 the detailed population structure of *S. flexneri* was understudied. Typically, subdivision of strains is through the antigenic variation of the O-antigen component, however, the subtyping methods lack accuracy and resolution contributing to the poor understanding of this species (Sun et al., 2011). Through the limited genomic studies done S. *flexneri* was observed as two distinct lineages. The first lineage (*S. flexneri* 6) consists of a single serotype and clusters within other *S. boydii* lineages (Choi et al., 2007). The second lineage is monophyletic and comprises of all the other *S. flexneri* serotypes and so is responsible for the majority of cases globally.

Connor et al. (2015) completed a phenomenal study to drive in depth understanding of the population structure and evolution of *S. flexneri* globally. A representative collection of *S. flexneri* strains were collected and WGS from key areas of endemic disease and spanning serotypes, 1-5, X, Xv and Y. A detailed modern population structure was determined which showed seven phylogenetically distinct phylogroups (PG) each containing isolates from different geographic locations and temporal dates separated by considerable evolutionary distances (Connor et al., 2015). Furthermore, Bayesian evolutionary analyses revealed that PGs 1, 2, 4 and 6 were the oldest lineages with most recent common ancestors (MRCA)

dating between 1341 and 1659 (Connor et al., 2015). Serotype 2a isolates, common vaccine targets, originated much later, 1822 onwards. It has been noted that that emergence of new *S. flexneri* PGs does not result in displacement of isolates of other PGS, instead the previous PGs persist and cause disease alongside the emerging PGs.

1.4.2. Serotypes

S. flexneri represents the *Shigella* species with the most diverse number and range of serotypes. As previously discussed, serotypes emerge through variation of the O-antigens either by glycosylation or O-acetylation of its sugar residues by phage encoding serotype converting genes. Currently, *S. flexneri* has 19 serotypes and counting (Nisa et al., 2020). In different geographical locations and temporal ranges, different serotypes are prevalent. In Asian countries, for example, the most prevalent serotypes are 2a, 1a and 3a with new serotypes and serotype switching emerging frequently to evade infection-induced immunity (Ye et al., 2010). Recently, in China, there has been the emergence of a novel serotype titled Yv, originating through a variant of the O antigen transferase gene *opt* (Sun et al., 2013). Serotype 2a has also been observed and is prevalent in India as well as outbreaks documented in U.S (Muthuirulandi Sethuvel et al., 2017, Reller et al., 2006).

1.4.3. Geography and burden

S. flexneri is endemic to many LMIC such as Bangladesh, India and Pakistan but has also been documented less frequently in higher income nations such as th U.S, typically associated with travel (Nisa et al., 2020). Typically, transmission of *S. flexneri* is through the faecal-oral route. In LMIC the lack of adequate hygiene facilities has led to *S. flexneri* being prevalent in paediatric populations. *Shigella* was the second leading cause of diarrhoeal mortality in 2016 and accounts for 14% of all diarrhoeal deaths in children under 5, representing a global public health burden (Khalil et al., 2018). Also contributing to *S.*

flexneri status as a global public health concern is the WHO recognising *Shigella* as a priority organism in terms of AMR (Shrivastava et al., 2018). Of particular concern is the increasing frequency of fluoroquinolone resistance in *S. flexneri*. Ciprofloxacin, a third-generation fluoroquinolone, is the primary antibiotic of choice for shigellosis, however, there have been reports of increasing ciprofloxacin resistance. The genetic mechanism(s) underlying resistance is commonly attributed to mutations in the quinolone resistance determining region (QRDR), ultimately diminishing the interaction between the antimicrobial and its target protein. In a study by Azmi et al. (2014) ciprofloxacin resistance increased from 0.7% in 2005 to 45.5% in 2010 in Bangladesh. The emergence of fluoroquinolone resistance through mutations in the QRDR as well as a mutation outside of the QRDR region in the *gyrA* gene has undermined current treatment strategies (Azmi et al., 2014).

Historically shigellosis has been associated with travel and only low levels of domestic acquisition in high-income nations, however, recently there have been numerous epidemics of domestically acquired shigellosis associated with transmission among adult males. Shigellosis has been reported as a sexually transmitted infection in men who have sex with men (MSM) populations since the 1970's and this type of transmission has become important for high income nations. Since 2009, there has been an increase of UK-acquired *S. flexneri* especially within the MSM population aged between 30 and 50 years, of which most belong to serotype 3a (Borg et al., 2012). Previously, the most prevalent serotype was 2a among MSM populations, however, there has been a shift to serotype 3a in recent years (Borg et al., 2012). In fact, the outbreak serotype 3a has been observed to have overtaken *S. sonnei* in Canadian MSM outbreaks and in the UK is now considered endemic. Furthermore, between 2008 and 2011 there was over a 400 % increase in the number of men diagnosed with *S. flexneri* 3a (Gilbart et al., 2015).

S. flexneri is the endemic *Shigella* species in LMIC and is associated with high rates of mortality and morbidity. The increasing prevalence of a plethora of AMR determinants is contributing to *S. flexneri's* reputation as a global public health concern. With *S. flexneri* transmission in MSM populations as well as through the typical faecal-oral and travel routes, studies into the evolution of *S. flexneri* within these scenarios could elucidate key factors involved in this adaptation.

1.5. An introduction to *S. sonnei*

S. sonnei is the most prevalent *Shigella* species within high-income nations and those that are undergoing economic development. However, also contributes to the global burden of disease within LMIC. In countries undergoing economic development, such as China in the last 50 years, the displacement of *S. flexneri* with *S. sonnei* is frequently observed. For example, a systematic review of shigellosis in mainland China between 2001 and 2010 revealed opposing trends with *S. flexneri* decreasing and *S. sonnei* increasing especially in the East, North and Northeast regions , a proposed association of regional economic development (Chang et al., 2012). *S. sonnei*, in contrast to *S. flexneri*, shows more limited genetic diversity and shares a singular serotype (Holt et al., 2012a).

1.5.1. Population structure

The global population structure of *S. sonnei* is split into five major lineages which can be further split into sublineages with isolates from similar geographical location but also through WGS studies emergence of publicly important sublineages have been observed and associated with the acquisition of AMR determinants (Holt et al., 2012a, Rabaa et al., 2016). Frequent in depth WGS studies been carried out to investigate *S. sonnei* epidemiology and detailed population structures at a regional level e.g Asia, Australia and South America as well as detailed population structures for specific traits e.g resistance, transmission in

orthodox Jewish communities (OJC) or MSM populations (Chung The et al., 2016, Baker et al., 2017, Ingle et al., 2019). This was causing confusion as there were now different names for the same lineage depending on which context you are viewing the population structure through. Even with the increasing number of in-depth WGS studies, which aid definition at the sublineage level, pre 2021 there had been a lack of a global genomic framework and genotype nomenclature impeding reporting and outbreak detection (Hawkey et al., 2021). Genotype frameworks based upon single nucleotide variants (SNVs) have been implemented in other bacterial species such as *Mycobacterium tuberculosis* (Coll et al., 2014). The main advantage of these genotype frameworks is that they allow fast typing without the need for complex comparative genomics.

Recently Hawkey et al. (2021) implemented a novel genotypic framework to *S. sonnei* WGS developed from 1935 isolates spanning 48 countries between 1943 and 2018. The successful genotyping scheme prioritised SNVs that are found in highly conserved core genes defining 137 genotypes (Hawkey et al., 2021). Additional high-resolution genotypes nested at the subclade level were added to differentiate between monophyletic groups based on AMR or transmission patterns and new genotypes can be added easily as they emerge (Hawkey et al., 2021). The genotyping framework from Hawkey et al. (2021) successfully utilised and released in an easily accessible format represents a universal nomenclature for *S. sonnei* especially important for tracking AMR *S. sonnei* clades at local and global levels.

1.5.2. Geography and burden

S. sonnei is prevalent in high-income countries and increasingly observed within economically developing nations where it displaces *S. flexneri* as the prominent species. The lack of *S. sonnei* in LMIC could be explained through passive environmental immunization

involving the Gram-negative bacteria *Plesiomonas shigelloides*. *P. shigelloides* serotype O17 and *S. sonnei* share an identical O-antigen structure (Kubler-Kielb et al., 2008). Due to the structural similarity of LPS between the two, vaccines prepared from the O-specific polysaccharides of *P. shigelloides* have been shown to be reasonably effective in preventing *S. sonnei* infection in humans (Kubler-Kielb et al., 2008). Also due to the cross-reactive nature of the O antigens it can be suggested that exposure to the *P. shigelloides* serotype O17 can lead to protection against *S. sonnei* infection (Sack et al., 1994).

In Asia, Europe, Australia and North America the most prominent *S. sonnei* Lineage for the past 20 years has been Lineage 3 (Holt et al., 2012a). The success of Lineage 3 strains globally is due to the acquisition of AMR determinants. Many Lineage 3 strains dominant in these countries are resistant to tetracycline and streptomycin, common first-line antibiotic treatments, which is acquired through a chromosomal class 2 integron, Tn7, and spA plasmid (Yang et al., 2005, McIver et al., 2002). Furthermore, there have been isolates with acquired macrolide and quinolone resistance determinants. Macrolide resistance was conferred through the resistance gene *mphA* and so led to the azithromycin resistance, an attractive treatment option for *Shigella* (Baker et al., 2017, Boumghar-Bourtchai et al., 2008). A particularly worrying example of resistance within Lineage 3 is through ciprofloxacin resistance, the recommended treatment for *Shigella* infections. Within Asia, ciprofloxacin resistant strains are being increasingly isolated forming a single clade in the expansion of Lineage 3 (Chung The et al., 2016). The resistance is conferred through mutations in the *gryA* and *parC* genes and can be globally attributed to a single clonal emergence event likely to have been established in South Asia (Chung The et al., 2016). Although ciprofloxacin resistance is more sporadic outside of Asia, strains exhibiting

ciprofloxacin resistance has been observed with increasing frequency in MSM populations in Canada and Asia (Gaudreau et al., 2011, Chiou et al., 2016).

S. sonnei is also highly endemic to Israel. *S. sonnei's* prevalent incidence rate is driven by biennial epidemics within the OJC in Israel (Baker et al., 2016). The OJC represents a risk group for *S. sonnei* and may be highly susceptible due to their densely populated living conditions, high numbers of young children and intracommunity transfer which can occur frequently due to social gatherings (Cohen et al., 2014). It is not just within Israel that the OJC represents a risk group but worldwide. In countries outside of Israel it has been shown that outbreak strains in OJCs are distinct from strains circulating in the general population. It was observed that OJC-associated strains were more closely affiliated with outbreaks associated with OJCs in other countries and strains circulating in Israel (Cohen et al., 2014). For example, in Antwerp, Belgium there is an OJC with approximately 10,000 persons living in one area of the town where in 2008 there was an *S. sonnei* outbreak (De Schrijver et al., 2011). Between April and August 2018 42 cases were recorded and all characterised isolates were found to share an identical profile and were indistinguishable from one of the twelve strains detected in Israel during 2008 (De Schrijver et al., 2011). This demonstrates the closeness of OJC associated outbreak strains.

Similar to *S. flexneri*, *S. sonnei* can be associated with MSM populations. For example, in Switzerland there has been the first report of sexually acquired *S. sonnei* which was MDR (Hinic et al., 2018). The genomic analyses demonstrated that the isolates fell into clusters from within UK MSM populations but with greater divergence. All of the isolates showed phenotypic resistance to azithromycin and two isolates showed resistance to quinolones (Hinic et al., 2018). Sexual transmission of *S. sonnei* causes increased prevalence in high-income countries where previously *Shigella* infections were infrequent.

S. sonnei represents an important *Shigella* species contributing to the global disease burden. In high income and economically developing nations *S. sonnei* is the prominent species where the infections contribute to pressure on healthcare systems and mortality rates.

1.6. Aims of research project

In this era of WGS and bioinformatic tools, understanding what factors have contributed to the long-term success of public health pathogens will significantly advance our ability to treat, manage and prevent these public pathogens. **In this thesis I aimed to utilise novel and traditional GWAS approaches to identify key factors contributing to the long-term success of *Shigella* species as global public pathogens.**

1.6.1. Overview of work

In chapter 2, I investigated factors positively associated with time that were contributing to the success of *S. sonnei* as a pathogen. First, I contextualised the *S. sonnei* historical isolates from the Murray Collection in terms of their place within the modern population structure, genotyping and their AMR and virulence profiles. Then through combination of historical *S. sonnei* isolates from the Murray collection and the plethora of readily available modern isolates, I was able to investigate the evolutionary arc of *S. sonnei* through a novel GWAS approach, herein coined temporal GWAS (tGWAS). I observed positive controls such as AMR and virulence factors to confirm the validity of tGWAS and then identified new insights into what other factors are contributing to the long-term success including iron metabolism genes and a plethora of catabolic mechanisms.

In chapter 3, I investigated factors positively associated with time that were contributing to the success of *S. flexneri* as a pathogen. First, I contextualised the *S. flexneri* historical isolates from the Murray Collection in terms of their place within the modern population structure, serotyping and their AMR and virulence profiles. Then through combination of

historical *S. flexneri* isolates from the Murray collection and the plethora of readily available modern isolates, I was able to investigate the evolutionary arc of *S. flexneri* through tGWAS. I observed positive controls such as AMR and virulence factors to confirm the validity of tGWAS and then identified new insights into what other factors are contributing to the long-term success including IS elements and the importance of the 'blackhole' of hypothetical genes. In depth analysis of the tGWAS results through SNP characterisations and Alphafold modelling to elucidate function of hypothetical proteins was undertaken. Identification of a previously hypothetical gene as a novel putative adhesin, adhesin Stv, was a truly exciting result. I identified adhesin Stv throughout key *S. flexneri* PGs and their expansions as well within the globally successful *S. sonnei* Lineage 3 and within other bacterial taxa. All suggesting the importance of this novel putative adhesin to the long-term success of a wide range of pathogens.

In chapter 4, I utilised traditional GWAS approaches to elucidate the contributing factors to the advantageous *E. coli* killing ability observed in *S. sonnei* Lineage 3 isolates. This chapter is a combination of laboratory work and bioinformatic approaches. GWAS revealed colicins to be a major contributing factor to the *E. coli* killing ability. In depth genotypic investigation revealed, however, that it was not a singular colicin responsible but a wide array of colicins contributing to this phenotype. Mass spectroscopy on killing isolates revealed the successful production of colicins by *S. sonnei* providing further evidence to support colicins as the factor contributing to this interbacterial competition advantage and hence aiding the long-term success of *S. sonnei* as a pathogen.

Chapter 2

2. Temporal GWAS identifies key factors contributing to the long-term success of *S. sonnei* as a pathogen

2.1. Introduction

Bacteria of the genus *Shigella* are a major contributor to the global diarrhoeal disease burden causing >200,000 deaths per annum globally where *S. flexneri* and *S. sonnei* are the major pathogenic species (Sahl et al., 2015, Khalil et al., 2018). Increasing antimicrobial resistance (AMR) in *Shigella* and the lack of a licenced vaccine has led WHO to recognise *Shigella* as a priority organism for the development of new antimicrobials (Shrivastava et al., 2018). Understanding what drives the long-term persistence and success of this pathogen is critical for ongoing shigellosis management and is relevant for other enteric bacteria.

S. sonnei causes significant disease in low- and middle-income countries (LMIC) and is the highest contributor to shigellosis in high-income nations or nations that are undergoing economic development (Hawkey et al., 2021). Within high-income nations shigellosis is frequently isolated from returning travellers or through men who have sex with men (MSM) (Baker et al., 2015c, Ingle et al., 2019). The global population of *S. sonnei* has been defined into five major lineages and share a single serotype (Holt et al., 2012b). *S. sonnei* whole genome sequencing (WGS) studies have been completed in multiple countries, including Asia (Holt et al., 2013, Chung The et al., 2016), Australia (Ingle et al., 2019) and the United Kingdom (Baker et al., 2018a), defining detailed population structure and sub-lineage groups of epidemiological importance – MSM communities or Orthodox Jewish communities (Baker et al., 2016, Rew et al., 2018, Ingle et al., 2019). Recently a genotypic framework of *S. sonnei* was created to provide a universal nomenclature enabling clear communication between research groups and public health officials (Hawkey et al., 2021). For the past 20 years Lineage 3 of *S.*

sonnei has dominated Europe, North America, and Australia likely because of the ongoing acquisition of key AMR determinants (Holt et al., 2012a).

Increasing AMR within *S. sonnei* has been a growing concern for public health authorities globally. The globally dominant Lineage 3 strains are typically resistant to early first-line antibiotics (trimethoprim, tetracycline, and streptomycin) due to horizontally acquired AMR genes, the spA plasmid and the Tn7-like transposon (Holt et al., 2012b). Further resistance to chloramphenicol and ampicillin is conferred via the *Shigella* resistance locus (SRL), a 16.7kb element borne on a 66-kb pathogenicity island (PAI) (Turner et al., 2001). Resistance or reduced susceptibility to fluoroquinolones and ciprofloxacin have also emerged through acquisition of point mutations in the quinolone resistance determining region (QRDR) of *gryA* and *parC* (Holt et al., 2013, Chung The et al., 2016). Acquisition of AMR determinants has already been shown to be vital for success of *Shigella* species, driving new epidemic strains and allowing accumulation of AMR determinants over time (Njamkepo et al., 2016a, Holt et al., 2012b, Baker et al., 2015a, Connor et al., 2015, Holt et al., 2013). In tandem with ongoing surveillance of rising AMR within *Shigella,* a deeper understanding of evolution within *Shigella* (and other bacterial taxa), would benefit efforts to address the AMR crisis.

To better investigate what is underpinning the success of *S. sonnei*, I utilised the unique and invaluable resource of historical isolates. Previously there were challenges to consider before utilising historical isolates, however, improvements in sequencing technologies and sampling strategies have allowed concerns such as low quantities of DNA and contamination to be overcome (Bennett and Baker, 2019). Historical isolates allow unique and rich insight into long-term trends in bacterial evolution spanning key time periods e.g introduction of antimicrobials, wars and human migration events (Bennett and Baker, 2019). The plethora of insights to be gained from historical isolates has already been demonstrated through

investigation of the first isolate accessioned by the National Collection of Type Cultures (NCTC) (Baker et al., 2014). NCTC1, was isolated in 1915 from a British Forces soldier and comparative genomics of NCTC1 with more modern reference strains revealed the streamlined acquisition of genetic material over time corresponding to the intuitively critical functions of AMR, virulence and immune evasion (Baker et al., 2014). There was, however, an island of unknown function retained in time indicating that the accessory genome was also acquiring unknown functions of similar importance (Baker et al., 2014). This study, using a single historical isolate, highlighted the potential knowledge gap around what drives the long-term success of bacterial pathogens which is particularly important for *Shigella* which contain many genes of unknown function (Sen and Verma, 2020).

NCTC1 truly demonstrated the utility of historical isolates. Expanding the quantity of historical isolates could facilitate a broader perspective (with greater statistical power) of how accessory genome changes over time has impacted pathogen success. A historical isolate collection coupled with modern isolates could be a potentially powerful combination to elucidate on pathogen success over time. One such historical isolates collection is the Murray Collection. The Murray Collection was amassed by the eminent microbiologist Professor Everitt George Dunne Murray compromising several hundred *Enterobacteriaceae* collected during the pre-antibiotic era (1917-1954) from a wide range of geographic areas (Baker et al., 2015a). The collection contains 683 bacterial strains, over 90 of which are *Shigella spp* (Baker et al., 2015a). This collection was WGS as a public resource for scientific research and the isolates are held by the NCTC. The utility of the collection has already been demonstrated in seminal studies that have elucidated the profile of pre-antibiotic era *Enterobacteriaceae* plasmids and the evolution of key phenotypes in *Klebsiella spp.* (Hughes and Datta, 1983, Wand et al., 2015a).

This study aimed to utilise the unique timescale provided by the Murray Collection combined with the plethora of modern WGS *S. sonnei* isolates publicly available to identify and interrogate what factors have contributed to the long-term success of *S. sonnei* over time. I developed a novel GWAS strategy, temporal GWAS (tGWAS), where genetic factors positively associated with time as a continuous outcome variable are explored as potential contributors to pathogen success. I utilise AMR and virulence determinants as positive controls to confirm the validity of tGWAS as a methodology. In my work, tGWAS analyses identified multiple AMR and virulence determinants as well as factors to aid full exploitation of the rich nutrient resources of the human host. The factors identified as positively associated with time all had intuitively beneficial roles which would aid pathogen success. I show tGWAS as a novel valid GWAS strategy that has implications for investigating pathogen evolution in broader bacteria taxa as well as identifying key factors of success for the globally important *S. sonnei,* and potentially relevant to treatment and management of other enteric pathogens.

2.2. Materials and methods

2.2.1. Whole genome sequencing of isolates

Three main datasets were collated for this study. The first contains *S. sonnei* historical isolates from the Murray Collection (n=22, 1937-1954) and can be accessed from the ENA databases project reference number PRJEB3255. These isolates were whole genome sequenced via Illumina HiSeq with 150 bp paired end reads. Secondly, a collection of isolates from Holt et al. (2012b) investigation into the population structure of *S. sonnei* were utilised (n=132). These isolates underwent sequencing via the Illumina Genome Analyzer GAII to generate tagged 54bp paired end reads. To complement, supplementary isolates were utilised from Baker et al. (2018a) (n=24). Thirdly, a representative subset of the Hawkey et al. (2021) dataset was used to complete the modern *S. sonnei* isolates utilised for

this study (n=127). These isolates were whole genome sequenced also using an Illumina Hiseq using 100 bp paired end reads. The reference strain *S. sonnei* strain 53G (HE616528.1) and its multiple plasmids (HE616529.1, HE616530.1, HE616531.1, HE616532.1) were employed for the duration of this study.

All raw sequence data was adapter- and quality- trimmed using Trimmomatic v0.38 (Bolger et al., 2014) and draft genomes were assembled using Unicycler v0.4.7 (Wick et al., 2017a). Annotation of genomes were completed using Prokka v1.13.3 (Seemann, 2014).

2.2.2. Phylogeny construction

Trimmed sequence (FASTQ) files were mapped against the reference strain *S. sonnei* strain 53G concatenated with its virulence plasmid (HE616528.1 – chromosome and HE616529.1, HE616530.1, HE616531.1, HE616532.1 – virulence plasmids) using Burrows-Wheeler Aligner (BWA) (Li and Durbin, 2009). The mapping files were filtered and sorted using samtools (Li et al., 2009a). Duplicates were marked using the tool Picard. Subsequent variant calling was completed through bcftools and a consensus file was generated for each isolate (Danecek and McCarthy, 2017). Each chromosome sequence was extracted, and regions were masked using a custom mask file containing plasmid sequences, IS elements and repeat regions. Gubbins was utilised to remove duplicate and low-quality sequences followed by SNP-sites to obtain the core-genome alignment (Page et al., 2016, Croucher et al., 2014). RAxML-ng was then utilised to infer a phylogenetic tree (Kozlov et al., 2019). Each phylogenetic tree has been midpoint rooted with visualisations completed using interactive Tree of Life (iTOL) v6.1.1 (Letunic and Bork, 2019).

2.2.3. Antimicrobial resistance and virulence determinants

The identification of genetic determinants contributing to AMR and virulence was completed using abricate v0.8.13 (Seemann). A minimum nucleotide identity threshold of

95% was used to ensure accurate identification of acquired resistance genes. For AMR the NCBI AMRFinder Plus database was utilised and for virulence the Virulence Factor Database (VFDB) was utilised (Feldgarden et al., 2019, Chen et al., 2015).

2.2.4. *Genotyping of* S. sonnei

Mykrobe v0.10.0 was utilised using the Mykrobe predict function on the FASTQ sequences of all isolates (Hunt et al., 2019a). The output from Mykrobe was then parsed using a custom python script (https://github.com/katholt/sonneityping) based on the genotyping scheme proposed by Hawkey et al. (2021). Upon parsing of the Mykrobe output a tsv file is generated containing the genotype.

2.2.5. Statistical testing

All statistical analyses were performed using R v3.6.1. To determine if there was a significant difference between the number of AMR determinants within the Murray Collection isolates and the modern isolate collection an independent t-test was undertaken. Similarly, to determine if there was a significant difference between the mean of the number of virulence factors within the Murray Collection isolates and the mean number of virulence factors within the modern isolate collection an independent t-test was undertaken

2.2.6. Genome Wide Association Studies

Paired end reads were mapped to the reference genome *S. sonnei* 53G using Burrows-Wheeler Alignment Tool (BWA) mem v07.17 (Li and Durbin, 2009) and Picard v2.23.1 was utilised to mark duplicates. Variant calling and subsequent filtering using Freebayes v1.3.2 (Garrison and Marth, 2012) was completed. The VCF files generated were merged and used as an input for the GWAS SNP analysis. To generate the input for the GWAS kmer analysis, kmers were counted from assemblies using fsm-lite v1.0. For the COG investigation, primary analysis was completed utilising the gene_presence_absence.Rtab file as input generated

from pan genome calculations via roary v1.007002 (Page et al., 2015b). Cross referencing of significant COGs was completed through a secondary COG GWAS analysis utilising the gene_presence_absence.Rtab file as input generated from pan genome calculations via panaroo v1.2.10 (Tonkin-Hill et al., 2020).

After generation of appropriate inputs GWAS was carried out using the tool Pyseer v1.3.6 (Lees et al., 2018). Pyseer uses linear models with fixed or mixed effects to estimate the effect of genetic variation in a bacterial population on a phenotype of interest, while accounting for potential confounding by population structure. For this investigation, time, in years, was utilised as the continuous phenotype. To account for population structure all analyses were supplemented with phylogenetic distances from the mid-point rooted core genome phylogeny already created as well as a covariate file containing lineages. Pyseer analyses were run using the linear mixed model (LMM).

To further explore key SNPs which resulted from GWAS, SnpEff v4.3.1 (Cingolani et al., 2012), a variant annotation and effect prediction tool was utilised. This tool was used to predict the functional effect of specific SNPs of interest.

2.3. Results and Discussion

2.3.1. Contextualisation of the Murray Collection isolates

2.3.1.1. Population structure

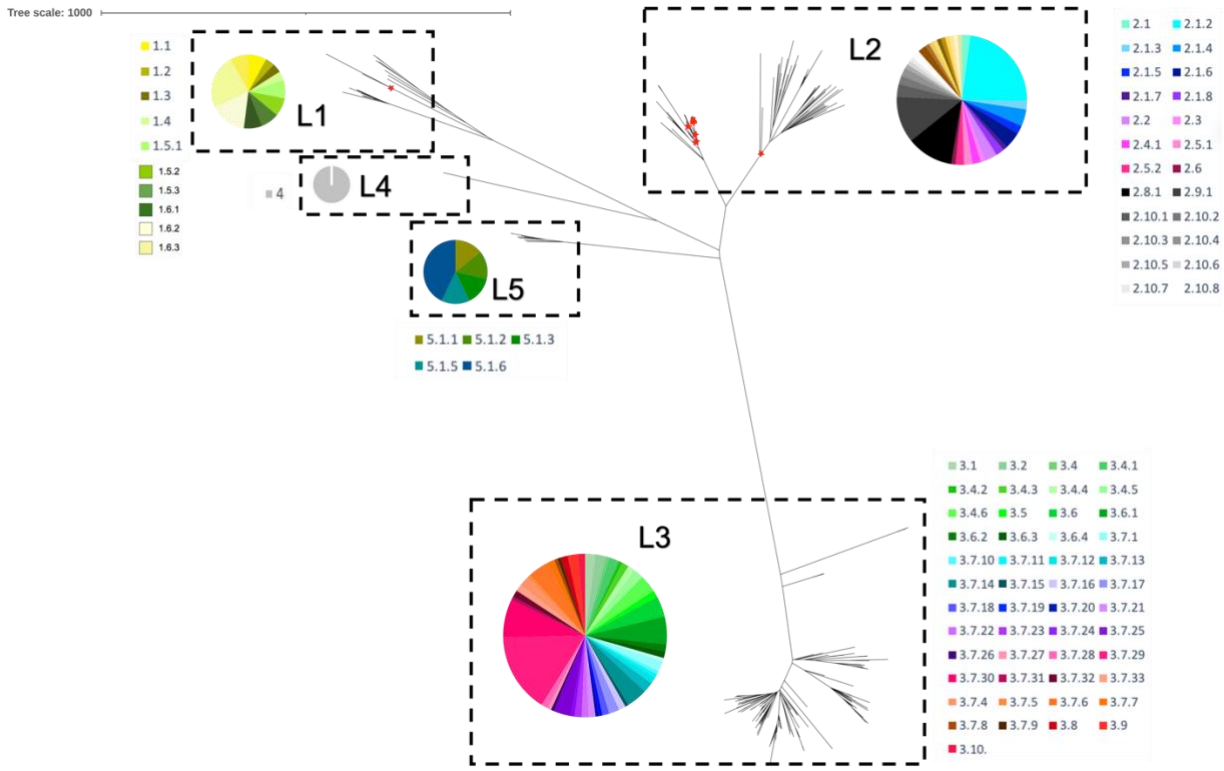


Figure 5: Contextualisation of historical isolates within the modern *S. sonnei* population structure.

Unrooted maximum likelihood phylogeny for *Shigella sonnei* isolates. Red circles indicate isolates which are from the Murray Collection. Lineages are clustered within the dashed boxes with the genotype composition within each lineage located in the corresponding pie chart. *S. sonnei* genotypes according to sonneityper are shown by colour in the pie chart according to the inlaid keys for each Lineage.

To determine whether and in what specific ways the population of the pre-antibiotic era *S. sonnei* differed from the modern *S. sonnei* population structure, Murray collection *S. sonnei* isolates were constructed into a phylogeny alongside modern *S. sonnei* isolates representing all five lineages. All sequence reads from the 242 isolates were mapped to the concatenated reference genome *S. sonnei* strain 53G with its three virulence plasmids to detect SNPs.

The Murray collection isolates were identified in Lineage 1 and Lineage 2 (Figure 5). Only a singular Murray isolate was identified in Lineage 1 isolated during 1937, the earliest year present in the *S. sonnei* Murray isolates. The other 21 Murray collection isolates were observed in Lineage 2, from 1937 to 1954. Identification of the *S. sonnei* historical isolates within Lineage 1 and 2 represents a unique insight into our understanding of the natural

history of S. sonnei. Newly emerging lineages of *S. sonnei* are typically observed displacing the older lineages and thus noting the historical isolates in the early lineages follows this biological understanding (Holt et al., 2012b). Previous temporal phylogenetic analyses estimated that the most recent common ancestor (MRCA) for Lineage 1 and 2 existed in the early 19th century whilst the MRCA for Lineage 3 existed at the turn of the 20th century (Holt et al., 2012b). All three of these lineages could encompass the pre-antibiotic era Murray isolates, however, observing the isolates within lineages 1 and 2 suggests their dominance during the pre-antibiotic era over Lineage 3. Lineage 3 of the *S. sonnei* species is considered to be the globally disseminating lineage whose success is thought to be due to the acquisition of key AMR determinants (Holt et al., 2012b). The proposed dominance of Lineages 1 and 2, could suggest that the selective pressure of antibiotic usage was needed for lineage 3 to obtain global dominance. However, there is a caveat to the proposed hypothesis. A significant proportion (n= 19/22, 86%) of the *S. sonnei* Murray isolates belong to a single year – 1937. The bias toward 1937 could offer a skewed representation of the pre-antibiotic era and not be representatively accurate of the time period. Thus Lineage 3 may be more prevalent than suggested by the phylogeny here and further analysis would need to be conducted in order to provide further evidentiary support for this hypothesis.

Recently, Hawkey et al. (2021) presented a genomic framework and genotyping scheme for *S. sonnei* which represents an exciting opportunity for universal nomenclature for *S. sonnei* allowing clear communication between public health and research groups (Hawkey et al., 2021). Genotyping of the S. sonnei Murray isolates was completed resulting in the historical isolates being assigned to genotypes 2,2.1, 2.1.2 and 2.2. Only a small number of Murray isolates were identified as genotypes 2 (n=1, 1937) and 2.1 (n=2, 1937). The majority of the

historical isolates (n=18/22, 82%) were assigned genotype 2.1.2 with isolates ranging from 1937-1954. There was a singular discrepancy between the lineage assignment from the phylogeny placement and the genotyping scheme. Specifically, an isolate from 1937 which was positioned within Lineage 1 of the phylogeny was assigned to genotype 2.2. The confidence in the genotype assignment was weak with a low-quality call for the final marker in the hierarchy, 2.2, where 66.9% of reads matching this marker belong to the alternative allele. Furthermore, there were several poorly supported markers for both Lineage 2 assignment and Lineage 2.2 assignment. It could be possible that there is an uncharacterised marker for Lineage 1 which is only present in historical isolates and no longer present in the modern isolates. This would mean the marker has not been incorporated into modern typing scheme and hence could result in the wrong genotype being assigned. Further investigation into the specific genotype markers within this singular isolate and the typing scheme would need to be completed to correct the discrepancy. This highlights the untapped genomic diversity that exists in historical isolate collections.

2.3.1.2. Increasing prevalence of AMR within *S. sonnei*

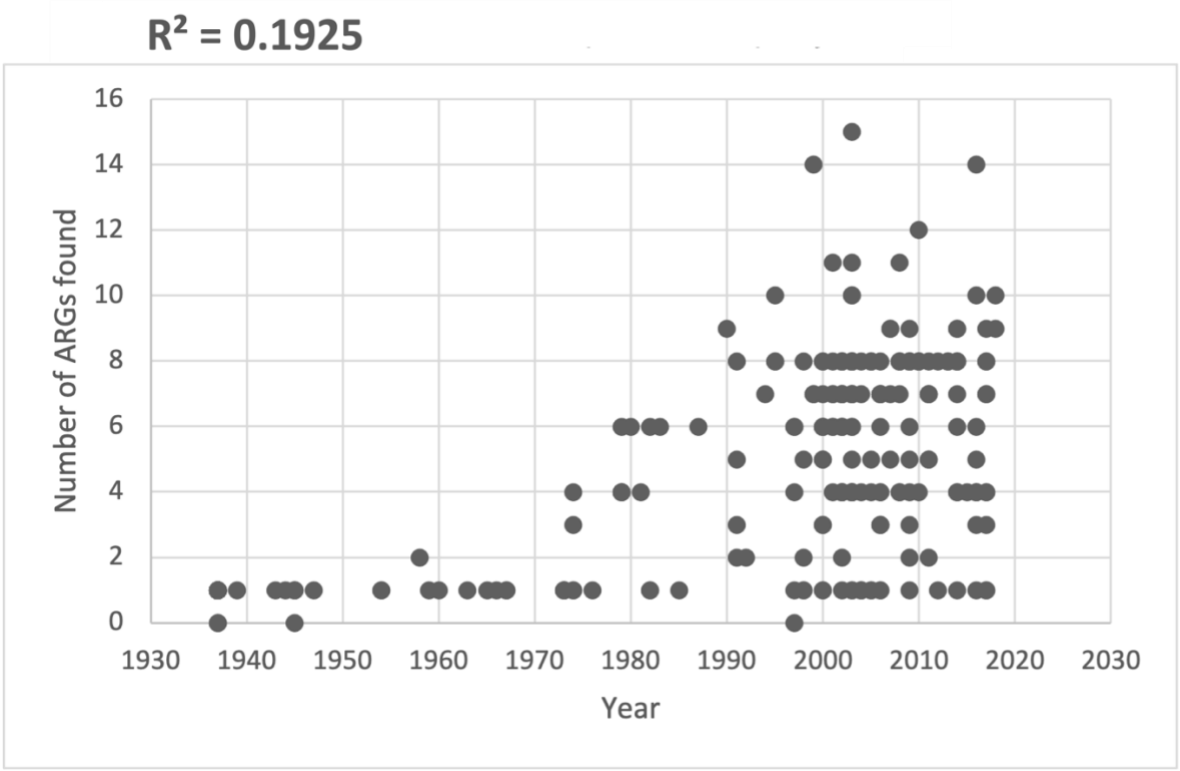


Figure 6: Temporal increase of AMR in *S. sonnei.*

The number of AMR determinants per isolate is shown over time, with a linear trend line shown and R^2 value noted at the top left.

Encompassing historical isolates from the pre-antibiotic era allows unique investigation into the evolution and acquisition of AMR determinants, highlighting how implementation of antibiotics has shaped bacterial genomes. In order to examine the evolution of AMR over time, genotypic AMR profiles were determined revealing that 91% of historical isolates (n=20/22) contained a single AMR determinant, *blaEC-8*. The *blaEC-8* gene confers penicillin resistance and is present with a wide range of *Enterobacteriaceae* with ancient origins (Kohler et al., 2022). The other two Murray *S. sonnei* isolates contained no AMR determinants and so genotypically would be predicted to confer no AMR resistance. Both of these isolates were isolated in 1937, representing the oldest year in this dataset, and

belonged to Lineage 2 clades, 2.1 and 2. The two isolates are genotypically in different clades and similarly this is exhibited within the phylogeny. It is plausible that due to the lack of early AMR determinants Lineage 2 was unable to establish global success and therefore died out and was outcompeted.

To explore the trend of AMR over the past 100 years within *S. sonnei*, genotypic AMR profiles were also determined for the modern collection. This revealed a positive correlation for the number of AMR determinants over time, with variation in the specific quantity of AMR determinants in each year (Figure 6). The variation in number of AMR determinants per isolate per year could be due to differing environmental conditions allowing varying rates of horizontal gene transfer and selective pressures. Statistical testing confirmed a statistically significant difference between the mean number of AMR determinants per isolate between the Murray (n=22, $\bar{x}$=0.91) and modern isolates (n=220, $\bar{x}$=5.21, t_{242} = 6.52, *P* <0.001). This increase in AMR determinants over time is consistent with previous descriptions in *S. dysenteriae* Type 1 and *S. flexneri c*over similar time frames (Njamkepo et al., 2016a, Connor et al., 2015). This investigation is consistent with broader trends in bacterial populations, and highlights that AMR is an appropriate marker for an accessory genome function that is increasing in *S. sonnei* within the time span of this dataset.

2.3.1.3. Overall virulence trends in *S. sonnei* are more difficult to elucidate

In a similar fashion to AMR, virulence factors are known to increase and be acquired over time, as demonstrated by NCTC1 (Baker et al., 2014). Similar to blaEC-8, there were a selection of virulence factors which were encoded in over 90% of Murray isolates and were conserved in the modern isolate collection. These factors (Table 1) had functions to do with iron uptake and various virulence functions such as host cell invasion, toxins and porin formation. All of these functions are intuitively beneficial for successful *S. sonnei* infection

where the organism must adapt to an iron deficient environment of the human body as well as invading mucosal cells via a plethora of virulence mechanisms. Two Murray isolates had 270% more virulence determinants than the rest (n=84 compared to n=31). These isolates were both isolated from 1937, representing the earliest year in the dataset, and are found in the same genotype, 2.1.2, and clade phylogenetically. The variation in quantity in virulence factors in these isolates compared to the other historical isolates, including isolates isolated from the same year, could be due to differing geographical locations and environmental conditions allowing greater gene transfer in bacterial communities or greater selective pressures guiding evolution or even an artefact of storage.

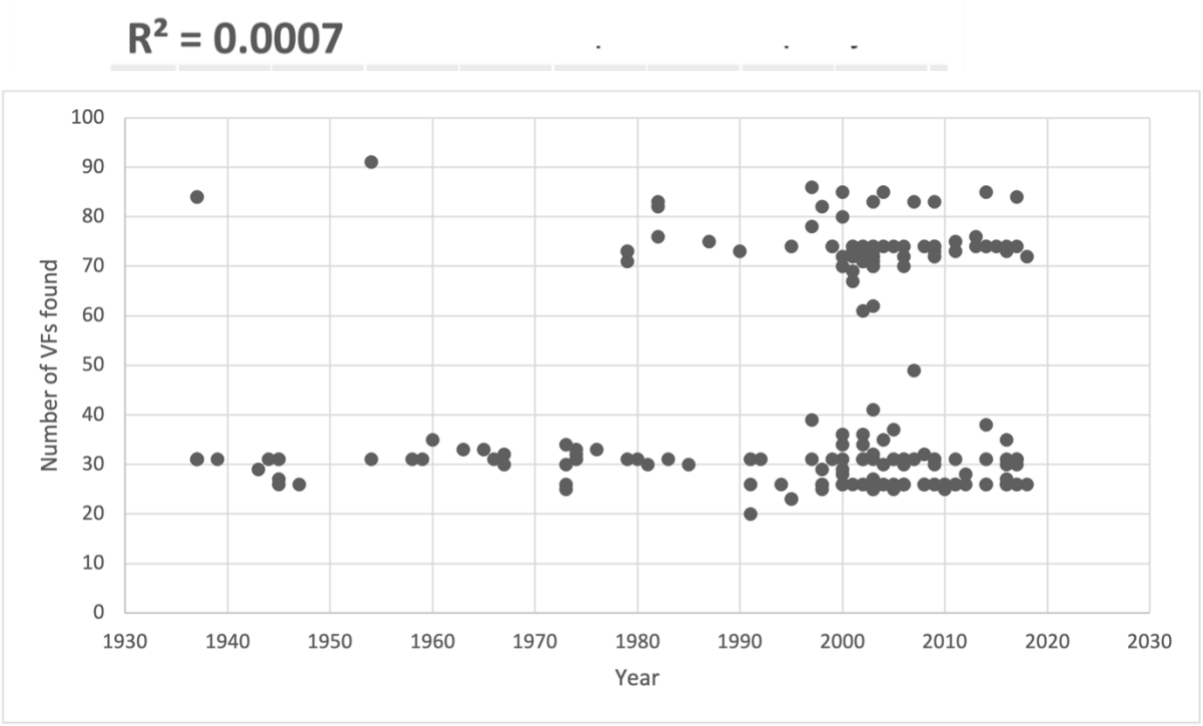


Figure 7: Temporal overview of virulence in *S. sonnei*.

The number of virulence factor (VF) determinants per isolate over time. The linear trend line is shown and R^2 value noted at the top left.

Overall trends for virulence factors (Figure 7) are more difficult to elucidate. There may be a tenuous positive correlation between the number of virulence factors per isolate over time, although there are high variation rates especially after 1980's, higher than those observed

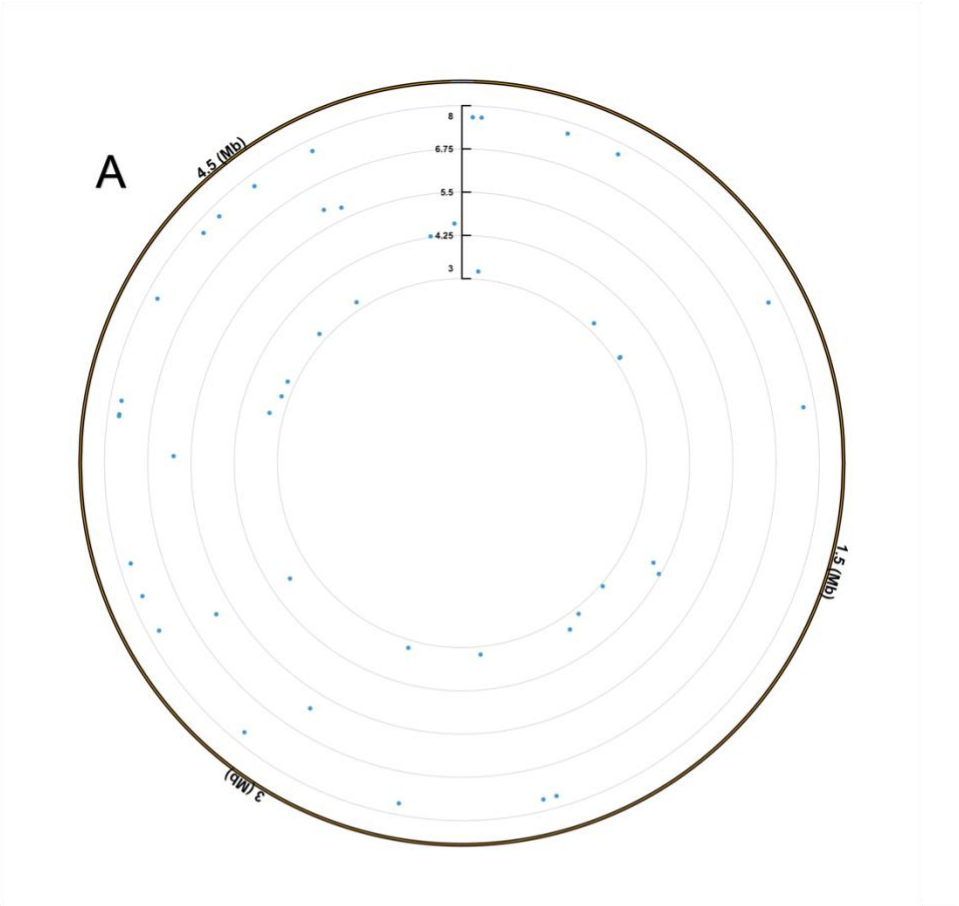
for AMR. Again, this variation is likely to be due to different environments with differing selective pressures and rates of horizontal gene transfer. Statistical testing, however, did show a statistically significant difference between the mean number of virulence determinants per isolate between the Murray (n=22, $\bar{x}$=36.3) and modern isolates (n=220, $\bar{x}$=43.6, t_{242} = 1.46, P <0.001). The statistical testing is consistent with broader trends in bacterial populations and highlights that virulence factors are an appropriate marker for an accessory genome function that is increasing in *S. sonnei* within the time span of this dataset.

Table 1: Virulence observed within historical *S. sonnei* isolates.

Table shows genes present in >90% of the Murray Collection *S. sonnei* isolates. Further detail in Supplementary Table 1.

Virulence genes found within pre-antibiotic era (Murray) *S. sonnei* isolates	**Function**	**Reference**
csgB/F/G	Host cell invasion	(Sakellaris et al., 2000)
entA/B/C/D/E/F/S	Iron uptake	(Wei and Murphy, 2016)
espL1,espX1,espX5	Regulation of cell cycle	(Faherty et al., 2010)
fepA/B/C/D/G	Iron uptake	(Wei and Murphy, 2016)
fes	Iron uptake	(Wyckoff et al., 2009)
fimB/F/G/H	Host cell invasion	(Klemm and Schembri, 2000)
iucB/C/D	Iron uptake	(Runyen-Janecky et al., 2003)
iutA	Iron Uptake	(Runyen-Janecky et al., 2003)
ompA	Porin/essential for conjugation/bacteriophage receptor	(Ambrosi et al., 2012)
senB	Enterotoxin	(Niyogi et al., 2004, Gu et al., 2019)
sigA	Serine Protease	(Al-Hasani et al., 2000)

2.3.2. An overview of the genetic factors associated with time via tGWAS


A
4.5 (Mb)
8
6.75
5.5
4.25
3
1.5 (Mb)
3 (Mb)

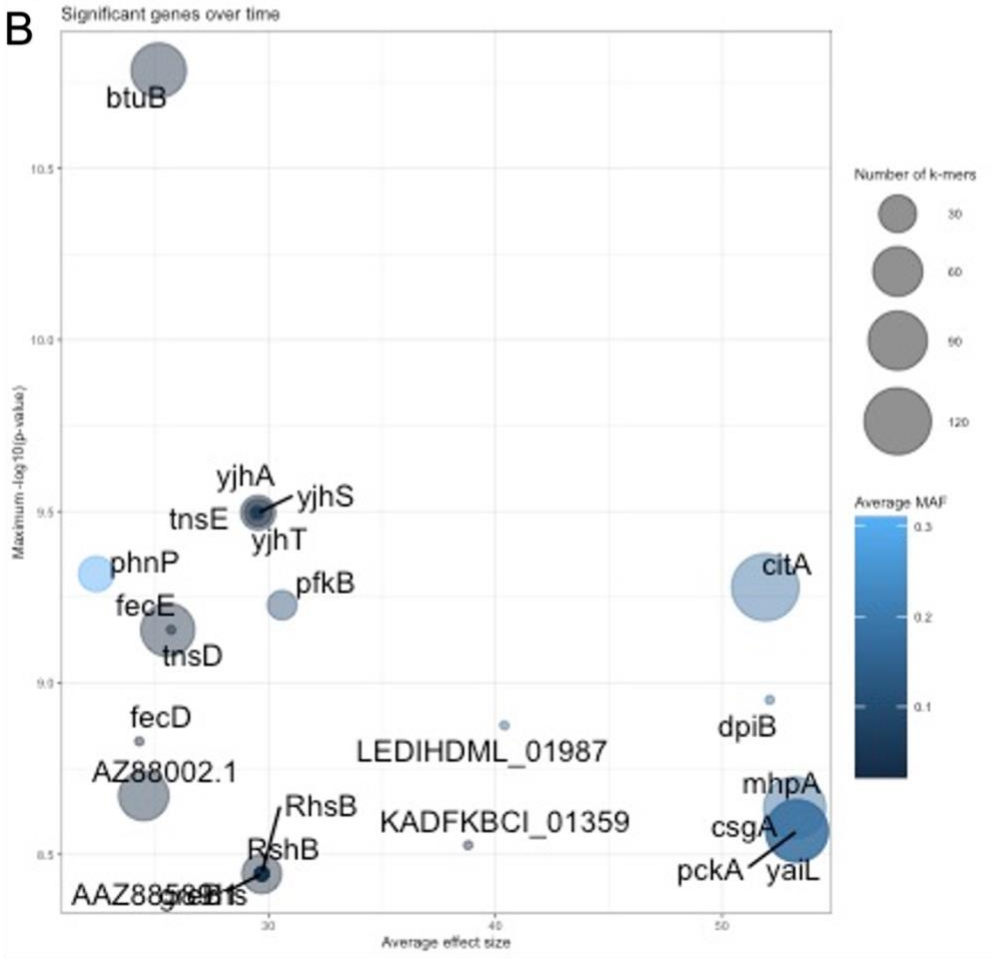

B
Significant genes over time
btuB
yjhA
yjhS
tnsE
yjhT
phnP
pfkB
fecE
tnsD
fecD
AZ88002.1
LEDIHDML_01987
RhsB
KADFKBCI_01359
RshB
citA
dpiB
mhpA
csgA
pckA
yaiL
Maximum -log10(p-value)
Average effect size
Number of k-mers
30
60
90
120
Average MAF
0.3
0.2
0.1

Figure 8: Genetic feature association in tGWAS by genetic feature type including SNPS (A) and kmers (B).

A SNPs: The circular manhattan plot shows the genomic position (clockwise) and negative logarithm of the p-value (radial axis) for 93 significantly positively associated SNPs **B kmers:** Bubble plot showing the number of kmers (bubble size) by gene (text labels in graph field), effect size (x-axis), and statistical support (y-axis as negative logarithm of the p-value).

Through conducting tGWAS various significant hits were observed for the three GWAS feature types. Specifically, SNPs as determined through read mapping, kmers (the presence of short genetic sequences of length k), and clustered orthologous groups (COGs) of predicted protein sequences. I identified a diverse range of SNPs (n=95, minlog10(p)>3, Figure 4), kmers (n=862, p-value<4.81E-09, Figure 4) and COGs (n=3181, lrt p-value<0.05) (Figure 8, Supplementary table 2, Table 2) that were significantly associated with the outcome variable of time of isolation (in years).

In order to focus on significant SNPs which would likely have an impact on biological function those SNPs which encoded for a missense mutation or stop codon within genes were focused on (n=45) (Figure 8, Table 2, Supplementary table 2). In total, 44 SNPs were located on the chromosome with only a singular missense SNP being located on one of the plasmids (HE616529.1). Observing no significant SNPs on the other two plasmids (HE616530.1 and HE616531.1) may indicate their stability as conserved mobile genetic elements and lower evolutionary rates.

In respect to the kmer and COG analyses, kmers of varying lengths (10-100 bp) were identified in 24 genes; 83% (n=20/24, Figure 8) of genes encoded for proteins with known functions and the remaining 17% (n=4/24) encoded for proteins with uncharacterised or hypothetical functions (Table 2, Supplementary table 2). COG analysis based on one method of COG clustering (roary) returned 3181 significant groups. Due to the vast quantity of COGs a second method of clustering (panaroo) was utilised and cross referencing to reduce COG numbers and strengthen their evidentiary support for downstream analysis. However, GWAS analysis via panaroo returned 3621 COGs and cross-referencing failed to reduce the

COGs by any significant amount. Thus, for practical reasons only the 25 most significantly associated hits (as measured by the lowest lrt p-values) from the roary clustering which were also present in the panaroo clustering were focused on for further investigation (Table 2, Supplementary Table 2). Most (88%, n=22/25) of the significant COGs encoded for proteins with known function and the remining COGs (n=3/25) were deemed hypothetical (Table 2). There were three genes in which multiple different GWAS types were detected and well supported as associated with time. These were *greB* and *fecD* which were found in significantly associated kmers and COGs whilst *trpGD* was found within significant SNPs and COGs (Table 2). The vast quantity of genetic factors significantly associated with *S. sonnei* over time offers a potentially unique insight into the underpinning components of the evolution and success of *S. sonnei* pathogens over time.

Table 2: Gene-associated genetic features associated with time in *S. sonnei*.

Genes and whether they were identified in SNP, kmer, or COG (25 most significantly associated) tGWAS analysis and related references are shown. ND – not described or investigated as part of this study. Additional details found in Footnote 1 and Supplementary Table 2.

Gene	SNP	Kmer	COG	Function	Reference
yaiL		√		Uncharacterised	ND
AAZ88002.1		√		Hypothetical	ND
pckA		√		Carboxylkinase involved in gluconeogenesis	(Medina et al., 1990)
btuB		√		Translocation of Vitamin B12	(Lundrigan et al., 1991)
greB		√	√	Transcription elongation factor	(Rutherford et al., 2007)
pfkB		√		phosphofructokinase involved in the first step of glycolysis	(Daldal, 1984)
csgA		√		Major Curlin subunit of the curli fimbriae	(Chanin et al., 2019a)
fecE		√		Iron transport	(Wei and Murphy, 2016)
citA		√		Citrate synthase	(Yang et al., 2005)
mhpA		√		Hydoxylase for Propionate catabolism	(Xu and Zhou, 2020)
phnP		√		Involved in phosphonate metabolism	(Hove-Jensen et al., 2011)
yjhT		√		Hypothetical	(Koestler et al., 2018)
tnsE		√		Transposon Tn7 transposition protein	(Peters and Craig, 2001)

yjhS		√		NanS in *E. coli*	(Steenbergen et al., 2009)
yjhA		√		Hypothetical	(Koestler et al., 2018)
AAZ88589.1		√		Hypothetical	ND
fecD		√	√	Iron Transport	(Wei and Murphy, 2016)
RhsB		√		Type IV protein	(Koskiniemi et al., 2013)
RhsB – another variant		√		Type IV protein	(Koskiniemi et al., 2013)
Rhs		√		Type IV protein	(Günther et al., 2022)
tnsD		√		Transposon Tn7 transposition protein	(Mitra et al., 2010)
fecA			√	Iron Transport	(Wei and Murphy, 2016)
fecB			√	Iron Transport	(Wei and Murphy, 2016)
trpE			√	Anthranilate synthase component 1	(Manson and Yanofsky, 1976)
trpGD	√		√	Involved in biosynthesis of anthranilate	(Manson and Yanofsky, 1976)
btsT			√	Pyruvate/proton symporter	(Kristoficova et al., 2018)
feoA			√	Iron Transport	(Runyen-Janecky et al., 2003)
feoC			√	Iron Transport	(Runyen-Janecky et al., 2003)
LEDIHDML_00428			√	YqeB	(Lin et al., 2015)
LEDIHDML_00010			√	Transporter	ND
LEDIHDML_03751			√	YbdD	ND
hslO			√	Molecular chaperone involved in bacterial defence to oxidative stress	ND
hslR			√	Heat Shock protein	ND
ompR			√	Transcription factor	(Bernardini et al., 1990)
rhaR_4			√	Transcriptional activator in l-rhamnose catabolism	(Tobin and Schleif, 1987b)
yjiA			√	Metal binding	(Sydor et al., 2013)
pdhR_2			√	pyruvate dehydrogenase complex regulator	(Feng and Cronan, 2014)
fimG			√	Involved in regulation of length and mediation of adhesion of type 1 fimbriae	(Chanin et al., 2019a)
gntP			√	Gluconate permease	(Klemm et al., 1996)
trpB			√	Tryptophan synthase beta chain	(Manson and Yanofsky, 1976)
cadC_2			√	Transcriptional regulatory protein of the cadmium resistance system	(Casalino et al., 2010)
LEDIHDML_02034			√	Hypothetical	ND
LEDIHDML_02035			√	Hypothetical	ND
LEDIHDML_02036			√	Hypothetical	ND

aaaT	√			ND	ND
aaeB	√			Acid efflux pump subunit	(Dyk et al., 2004)
amiA	√			Involved in septum cleavage during cell division	ND
caiA	√			Crotonobetainyl-CoA reductase	ND
chrR	√			Catalyses the reduction of quinones	(Eswaramoorthy et al., 2012)
cscB	√			Transport of sucrose into the cell	(Sahintoth et al., 1995)
cutC	√			Contributes to copper tolerance	(Gupta et al., 1995)
dgcE	√			Involved in the control of the switch from cell motility to adhesion	(Malekian et al., 2022)
dnaB	√			initiates replicative DNA synthesis	ND
ecfA1	√			TP-binding (A) component of a common energy-coupling factor (ECF) ABC-transporter complex	ND
ehaG	√			Trimeric autotransporter proteins	(Totsika et al., 2012)
eptB	√			Catalyse the transfer of phosphoethanolamine to lipid A	(Elizabeth et al., 2022)
fadB	√			Degradation of fatty acids	(Campbell et al., 2003)
fadH	√			Fatty acid degradation	(Pavoncello et al., 2022)
fdhF	√			Decomposes formic acid	ND
frlD	√			Catalyzes the ATP-dependent phosphorylation of fructoselysine	ND
gatY	√			galactitol catabolism	(Nobelmann and Lengeler, 1996)
IBDECAPI_00251	√			Hypothetical	ND
IBDECAPI_00793	√			Hypothetical	ND
IBDECAPI_00795	√			Hypothetical	ND
IBDECAPI_01121	√			Hypothetical	ND
IBDECAPI_01885	√			Hypothetical	ND
IBDECAPI_02069	√			Hypothetical	ND
IBDECAPI_03070	√			Hypothetical	ND
IBDECAPI_04197	√			Hypothetical	ND
IBDECAPI_04536	√			Hypothetical	ND
IBDECAPI_05056	√			Hypothetical	ND
FJMHIPND_00026	√			IS3 transposase	(Hawkey et al., 2020)
IBDECAPI_01724	√			IS4 family transposase	(Hawkey et al., 2020)
leuA	√			2-isopropylmalate synthase	ND
norR	√			Required for the expression of anaerobic nitric oxide (NO) reductase	(Tucker et al., 2006)
nrfG	√			ND	ND
plsB	√			Glycerol-3-phosphate acyltransferase	ND
IBDECAPI_02538	√			Putative autotransporter	ND

rapA	√			RNA polymerase-associated protein RapA	(Lynch et al., 2007)
IBDECAPI_04696	√			Regulator of Sigma D	ND
rhaS	√			Rhamnose regulator	(Tobin and Schleif, 1987a)
tatB	√			Part of the twin-arginine translocation (Tat)	(Sargent et al., 1999)
ugpC	√			Part of the ABC transporter complex UgpABCE	ND
yaiY	√			ND	ND
ybeM	√			ND	ND
ybjI	√			Putative transport protein	ND
ygfX	√			ND	ND
yhhJ[1]	√			Transport permease	ND

2.3.2.1. AMR and virulence determinants detected in tGWAS

Upon examination of the full plethora of genetic features significantly associated with *S. sonnei* over time, AMR and virulence determinants were identified acting as a positive control for the validity of tGWAS. A singular SNP within the *eptB* gene was found (Table 2). *eptB* encodes for a Ca^{2+} induced phosphoethanolamine (PEtN) modifying Lipid A of the lipopolysaccharide (LPS) of bacteria leading to colistin resistance (Elizabeth et al., 2022). Colistin resistance is currently considered a serious problem due to its vital efficacy against most multi-drug Gram-negative bacteria and the rapidly increasing prevalence of colistin resistance in *Enterobacteriaceae* (Aghapour et al., 2019). Mutations in AMR genes can lead to heightened resistance and even a singular SNP such as the one identified through tGWAS could have a profound impact on *eptB* efficacy.

In addition to AMR several key virulence determinants were identified – kmers within *RhsB* and *rhs* genes as well as *ompR* found in the COG analyses (Table 2). Rhs and related proteins are widely distributed in both bacteria and in eukaryotes but have been shown to mediate

[1] The known gene names are taken from the annotated *S. flexneri* 2a strain 301 reference genome and unknown genes entitled SF or CP were also named from the reference genome and its virulence plasmid. Other unknown gene names were assigned based on the first genome they were observed in from the isolate collection.

intercellular competition (Koskiniemi et al., 2013). *RhsB* in other Gram-negative bacteria has been linked to type VI secretion systems (T6SS) where it thought to be exported as an effector protein (Koskiniemi et al., 2013). T6SS are employed by multiple Gram-negative bacterial species to inject effector proteins in a contact-dependent manner to neighbouring bacterial and eukaryotic cells (Monjarás Feria and Valvano, 2020). Rhs and related proteins share C-terminal toxin domains with WapA proteins from Gram-positive bacteria known to inhibit the growth of neighbouring cells (Monjarás Feria and Valvano, 2020). Hence the potential Rhs and related proteins may have a vital role for virulence within contact dependent growth inhibition. Similarly, *ompR* encodes for a key virulence factor. *ompR* encodes for a response regulator of the two-component signal transduction pathway EnvZ/OmpR responsible for invasion of epithelial cells, a key part of *Shigella* infection (Brzóstkowska et al., 2012, Bernardini et al., 1990).

Shigella's ability to adhere to human cells complements its invasion capabilities. The gene *fimG* was observed during the COG analysis (Table 2). Type 1 fimbriae are filamentous structures produced by multiple *Enterobacteriaceae* and encoded for by the *fimAICDFGH* operon (Bravo et al., 2015, Chanin et al., 2019a). *fimG* encodes for vital adaptor proteins which contribute to the fimbriae capacity to enhance adhesiveness and [i]invasiveness of *Shigella* species (Bravo et al., 2015).

Genes conferring AMR or encoding for virulence have key and intuitively beneficial roles for the success and visibility of *Shigella* as a pathogen. Quantitatively there were more virulence factors significantly associated with time than AMR related genes, this could indicate the increasing importance of virulence in an intracellular pathogen's evolution. tGWAS has successfully highlighted gene products key for AMR and virulence over time, indicating tGWAS validity as a novel methodology.

2.3.2.2. Nutrient mining and catabolic mechanisms detected by tGWAS

Shigella is an intracellular human host pathogen. Adapting successfully to the host environment involves overcoming multiple challenges including bacterial competition for nutrients from host microbiota and adapting to potentially different primary metabolites. Out of the 91 genes in which significant GWAS feature types fell, 7% (6/91) all shared iron uptake as a function (Table 2). Iron uptake has been previously highlighted in this study for its importance to the success of *S. sonnei* due to its high prevalence in the historical isolates (n=5/9) (Table 1). Once *S. sonnei* have successfully invaded the epithelial cells, it is essential that *S. sonnei* acquire the nutrients required for proliferation and dissemination. However, the human host is an extremely iron limiting environment, and it is essential that *S. sonnei* not only has mechanisms to obtain iron but outcompete other bacteria in the human microbiota also vying for the same source. The *fec* genes (*fecABED*) and *feo* genes (*feoAC*) (Table 2) are involved in iron acquisition systems specifically a ferric dicitrate uptake system and a ferrous iron uptake system respectively (Luck et al., 2001, Wei and Murphy, 2016, Runyen-Janecky et al., 2003). *fecAB* and *feoAC* were observed during the top 25 significantly associated tGWAS COG analysis whilst *fecE* contained 72 significant kmers (k=11-100bp) and *fecD* contained a singular significant kmer (k=12) and was observed in the COG tGWAS analyses (Table 2). Conservation of iron uptake systems over time as well as their appearance in multiple tGWAS feature types highlights their unceasing importance in the success of *S. sonnei* as an intracellular pathogen.

Throughout all GWAS feature types, multiple genes encoded for proteins involved in catabolism or synthesis of multiple different metabolites (Table 2). Significant kmers were observed within metabolism of citrate (*citA*, n=121 k=12-100bp), proprionate (*mhpA*, n=100, k=100bp) and phosphonate (*phnP*, n=26, k=100bp). All of these genes encode for proteins

involved in catabolism of citrate, propionate and phosphonate respectively (Yang et al., 2005, Xu and Zhou, 2020, Hove-Jensen et al., 2011). Furthermore, significant singular SNPs were observed in *trpGD*, *fadB*, *fadH* and *gatY* (Table 2) involved in anthranilate synthesis, fatty acid catabolism and galactitol catabolism respectively (Manson and Yanofsky, 1976, Campbell et al., 2003, Pavoncello et al., 2022, Nobelmann and Lengeler, 1996). *trpGD* was also noted in the top 25 COGs along with *trpE* further highlighting the importance of fatty acid degradation in *S. sonnei* over time (Table 2). The quantity of genes contributing to the wide range of catabolic mechanisms indicates their importance over time and highlights that part of *Shigella's* colonization success is potentially by fully exploiting the rich source of nutrients within the human host (Passalacqua et al., 2016).

tGWAS has identified multiple genes significantly associated with *S. sonnei* over time encoding for functions related to exploiting the rich human host nutrient resources. These genes encode for intuitively beneficial functions which would aid *S. sonnei's* ability to survive, colonize and replicate enabling successful *Shigella* infection within a human host.

2.4. Conclusion

This study has successfully contextualised the historical Murray Collection *S. sonnei* isolates within the modern population structure confirming previous biological understanding of linage displacement by emerging lineages. Concordant with evolutionary understanding of AMR, very few or no AMR determinants were present in the historical isolates representing the pre-antibiotic era. Early virulence determinants were identified and were of clear importance displayed by their conservation in the modern isolates. Iron uptake was the most common virulence agent historically (n=5/9) indicating its early importance to an intracellular pathogen. The broader biological trends of increasing AMR and virulence have also been validated, highlighting their potential to be used as positive controls for factors

positively associated with time. Successful tGWAS returned multiple AMR and virulence determinants indicating the validity of tGWAS as a strategy for identifying key factors positively associated with time. Other key factors identified via tGWAS had intuitively beneficial roles for the success of *S. sonnei* as a pathogen. Key genes for iron uptake and those encoding for a plethora of catabolic mechanisms were identified, highlighting their potential importance for *S. sonnei's* success.

2.5. Next steps

The plethora of results from tGWAS and their intuitively beneficial roles is promising for the validity of tGWAS as a methodology. However, the bias of the historical isolate data to the year 1937 means that the timespan of the pre-antibiotic era is not fully represented and so this may have slightly negatively skewed the results positively associated with time, leaving out key genes and functions. Due to this potential bias, it was decided that the *S. flexneri* dataset displayed a more representative picture of the pre-antibiotic era timespan and so further investigation of tGWAS hits would be focused on from *S. flexneri* analyses and not *S. sonnei*.

Chapter 3

Much of the content of this chapter was published in the research article "Temporal GWAS identifies a widely distributed putative adhesin contributing to pathogen success in *Shigella* spp" (Bennett et al., 2022) which is currently under review. Addition of section 3.2.10 and 3.3.9 were included for this thesis but not present in the manuscript. Permission to include the publication in this PhD was obtained from all co-authors. Specifically, I acknowledge the contribution of authors below. Unless specified below all work was completed by myself.

P. Malaka De Silva	Completed the confirmation that adhesin Stv was present on pStv via PCR
Malcolm J Horsburgh	Supervision
Tim R Blower	Supervision
Kate S Baker	Supervision

3. Temporal GWAS identifies a widely distributed putative adhesin contributing to pathogen success in *Shigella* spp.

3.1. Introduction

Bacteria of the genus *Shigella* are non-motile Gram-negative facultative anaerobic bacilli which cause shigellosis. Shigellosis is an intestinal infection characterised by fever, nausea, dehydration, and bloody diarrhoea, and is the second leading cause of diarrhoeal mortality causing approximately 163 million cases and >200,000 deaths annually (Sahl et al., 2015, Khalil et al., 2018). Like many other diarrhoeal pathogens, most disease occurs in low to middle-income countries (LMIC) in children under the age of 5 years, in which the majority (65%) of disease is caused by the species *S. flexneri* (Kotloff et al., 2013, Livio et al., 2014). *S. flexneri* consists of 7 different genomic phylogroups and 18 serotypes, as determined by the somatic (O) antigen located on the cell surface (Connor et al., 2015). Serodiversity is determined by acquisition of mobile genetic elements that confer or modify O-antigen types

(Connor et al., 2015, Allison and Verma, 2000, Liu et al., 2008a). The plasticity of the accessory genome of shigellae is further facilitated by the presence of hundreds of insertion sequences throughout their genomes and evidenced by the ability to acquire plasmids and integrative conjugative elements conferring traits such as antimicrobial resistance (AMR) (Hawkey et al., 2020, Njamkepo et al., 2016a).

Latterly, increasing AMR in *Shigella* has led to *S. sonnei* being recognised as a priority AMR pathogen by the World Health Organisation (WHO) (Shrivastava et al., 2018). A recent Organisation for Economic Co-operation and Development (OECD) report estimates that 2.4 million people in Europe, North America and Australia will die from resistant infections in the next 30 years at a cost ≥US$3.5 billion per year (OECD, 2018). With stark warnings of a post-antibiotic era where common infections could once treatable infections could again kill, it is evident that understanding the evolution of AMR and factors that accompany it are vital to overcome this global issue. Acquisition of AMR determinants has already been shown to be vital for success of *Shigella* species, driving new epidemic strains and allowing accumulation of AMR determinants over time (Njamkepo et al., 2016a, Holt et al., 2012b, Baker et al., 2015a, Connor et al., 2015, Holt et al., 2013). In tandem with ongoing surveillance of rising AMR within *Shigella,* a deeper understanding of evolution within *Shigella* (and other organisms), would benefit efforts to address the AMR crisis.

Historical isolate collections provide an invaluable resource with which to make unique insights into long-term trends in pathogen evolution and epidemiology, and evaluate the impact of major historical events, e.g. the clinical introduction of antimicrobials, on bacterial evolution (Bennett and Baker, 2019). Advances in both sequencing technologies and sampling strategies have allowed previously inaccessible historical bacterial DNA to be investigated, revealing a plethora of unique insights, such as those gained from the first isolate accessioned

into the National Collection of Type Cultures (NCTC) (Baker et al., 2014). This *S. flexneri* isolate, NCTC1, was isolated in 1915 from a British Forces soldier and comparative genomics of NCTC1 with more modern reference strains revealed the streamlined acquisition of genetic material over time corresponding to the intuitively critical functions of AMR, virulence and immune evasion (Baker et al., 2014). There was, however, an island of unknown function retained in time indicating that the accessory genome was also acquiring unknown functions of similar importance (Baker et al., 2014). This study, using a single historical isolate, highlighted the potential knowledge gap around what drives the long-term success of bacterial pathogens which is particularly important for *Shigella* which contain many genes of unknown function (Sen and Verma, 2020).

A broader perspective on how the accessory genome changes over time to facilitate pathogen success could be obtained by examining historical and modern isolate collections over centuries. One such historical collection of isolates is the Murray Collection, compromising several hundred *Enterobacteriaceae* collected during the pre-antibiotic era (1917-1954) from a wide range of geographic areas (Baker et al., 2015a). The collection was amassed by the eminent microbiologist Professor Everitt George Dunne Murray and comprises 683 bacterial strains, over 90 of which are *Shigella spp* (Baker et al., 2015a). This collection was whole genome sequenced (WGS) as a public resource for scientific research and the isolates are held by the NCTC. The utility of the collection has already been demonstrated in seminal studies that have elucidated the profile of pre-antibiotic era *Enterobacteriaceae* plasmids and the evolution of key phenotypes in *Klebsiella spp.* (Hughes and Datta, 1983, Wand et al., 2015a).

Here I expanded the scale and sophistication of looking at accessory genome changes over time by using the unique genomic resource of the Murray Collection complemented with more modern isolates to further characterise the factors that have contributed to the long-

term success of *Shigella* over time. I describe herein our novel strategy, temporal genome wide association study (tGWAS), where genetic factors positively associated with time as a continuous outcome variable are explored as potential contributors to pathogen success. I demonstrate the efficacy of this approach through the recovery of factors known to contribute to the critical functions of AMR and virulence, as well as unknown determinants, in *S. flexneri*. I developed a framework for prioritising exploration of tGWAS hits that had no assigned function, useful for the broader translation of tGWAS to other bacterial pathogens, through which I identified an undescribed but widely distributed putative adhesin; which I called Stv. In addition to being associated with clonal expansions in *S. flexneri*, I demonstrate the external validity of our findings using a global *S. sonnei* dataset that shows that Stv acquisition predated AMR development and global dissemination of the dominant *S. sonnei* subtype (Lineage III), supporting its role in driving pathogen success.

3.2. Materials and methods

3.2.1. Sequence data and basic processing

Two main publicly available *S. flexneri* datasets were collated for this study. The first contained WGS data from historical isolates of *S. flexneri* isolates from the Murray Collection (n=45, isolated between 1917-1935) and can be accessed from ENA project number PM463261343GB (Baker et al., 2015a). Genome sequence data from NCTC1 (GCA_000953045.1) was also incorporated into this study (included in the Murray collection during statistical analyses)(Baker et al., 2014). Secondly, a collection of isolates from a global study of *S. flexneri* was used as a modern isolate resource (n=262, isolated between 1950-2011) (Connor et al., 2015). The *S. flexneri* 2a strain 301 and its virulence plasmid (NC_004337.2 – chromosome and NC_004851.1 – pCP301 plasmid) were employed as a

reference strain. Individual isolate accession numbers, data, and metadata can be found in Supplementary Table 3.

To explore the presence of novel adhesin Stv in *S. sonnei*, three datasets of *S. sonnei* isolates were collated (Supplementary table 2) and utilised. The first contained *S. sonnei* historical isolates from the Murray Collection (n=22, 1937-1954) accessible from PRJEB3255 (Baker et al., 2015a); second, a collection of isolates (n=132) from an investigation into the global population structure of *S. sonnei* (Holt et al., 2012b); and supplementary isolates (n=24) from Baker et al. (2018a); and thirdly, a representative subset (n=127) of clade assignments, from Hawkey et al. (2021), were used as the modern *S. sonnei* collection used for this study (n=127). The reference strain *S. sonnei* strain 53G (HE616528.1) and its multiple plasmids (HE616529.1, HE616530.1, HE616531.1, HE616532.1) were employed as the reference sequence. Individual isolate accession numbers, data, and metadata can be found in Supplementary Table 1.

Raw sequence data was adapter- and quality- trimmed using Trimmomatic v0.38 (Bolger et al., 2014) and draft genomes were assembled using Unicycler v0.4.7 (Wick et al., 2017a) and annotated using Prokka v1.13.3 (Seemann, 2014).

3.2.2. Phylogeny construction

Trimmed sequence (FASTQ) files were mapped against the relevant reference strain using Burrows-Wheeler Aligner (BWA) v0.5.9-r16 (Li and Durbin, 2009). The mapping files were filtered and sorted using samtools v1.13 (Li et al., 2009a). Duplicates were marked using Picard (2019a) and variant calling was completed with bcftools to generate a consensus genome sequence for each isolate (Danecek and McCarthy, 2017). Each chromosome sequence was extracted, and regions were masked using a custom mask file (one for *S.*

flexneri and one for *S. sonnei*) containing plasmid sequences, IS elements and repeat regions (identified from reference genomes). Gubbins v2.3.4 removed duplicate and low-quality sequences followed by SNP-sites to obtain the core-genome alignment (Page et al., 2016, Croucher et al., 2014). RAxML-ng was then utilised to infer a phylogenetic tree (Kozlov et al., 2019). Each phylogenetic tree has been midpoint rooted with visualisations completed using interactive Tree of Life (iTOL) v6.1.1 (Letunic and Bork, 2019).

3.2.3. Antimicrobial resistance and virulence determinants

The identification of genetic determinants attributing to AMR and virulence was completed using abricate v0.8.13 (Seemann) using the NCBI AMRFinder Plus database and Virulence Factor Database (VFDB) (Feldgarden et al., 2019, Chen et al., 2015). A minimum identity threshold of 95 was used to ensure accurate identification.

3.2.4. Statistical testing

All statistical analyses were performed using R v3.6.1 (Team, 2013). To determine if there was a significant difference between the number of AMR determinants within the *S. flexneri* Murray Collection isolates and the *S. flexneri* modern isolate collection an independent t-test was undertaken and conducted similarly for virulence factors.

3.2.5. Temporal Genome Wide Association Studies

S. flexneri paired end reads were mapped to the *S. flexneri* 2a strain 301 reference genome using Burrows-Wheeler Alignment Tool (BWA) mem v07.17 (Li and Durbin, 2009) and Picard v2.23.1 was utilised to mark duplicates. Variant calling and subsequent filtering using Freebayes v1.3.2 was completed (Garrison and Marth, 2012). The VCF files generated were merged and used as an input for the GWAS single nucleotide polymorphisms (SNP) analysis. To generate the input for the GWAS kmer analysis, kmers (small substrings of nucleotides of varying length=k to account for short and longer variation) were counted from assemblies

using fsm-lite v1.0. For investigation of Clusters of Orthologous Genes (COGs), primary clustering was delivered as the gene_presence_absence.Rtab file from pan genome calculations via roary v1.007002 (Page et al., 2015b). Cross referencing of significant COGs with a secondary COG GWAS analysis was conducted utilising the gene_presence_absence.Rtab file generated from pan genome calculations from panaroo v1.2.10 (Tonkin-Hill et al., 2020).

After generation of appropriate inputs, GWAS was conducted using Pyseer v1.3.6 which uses linear models with fixed or mixed effects to estimate the effect of genetic variation in a bacterial population on a variable of interest while accounting for confounding population structure (Lees et al., 2018). For this investigation, time, in years, was used as the continuous phenotype. To account for population structure, all analyses were supplemented with phylogenetic distances from the mid-point rooted phylogeny and a covariate file of assigned phylogroups. Pyseer analyses were run using the linear mixed model (LMM). To further explore GWAS SNP hits, SnpEff v4.3.1 was used to predict the functional effect of annotated variants (Cingolani et al., 2012).

3.2.6. Prioritisation strategy for evaluating GWAS hits

Due to the plethora of genetic features (i.e. SNPs, kmers, and COGS) that were identified as positively associated with time in the tGWAS analyses, a decision tree was developed to guide which features to pursue (Supplementary Figure 1). The decision tree was designed to account for each genetic feature type that underpinned the three GWAS analyses. Specifically, these were: SNPs, kmers and COGs, with the aim of selecting the feature likely had important biological functions. The decision tree was based on varied criteria that differed by genetic feature type but involved consideration of whether hits related to genes of known or unknown function; the distribution of genetic features across the population

structure; the predicted function effect (synonymous, missense or nonsense SNPs); plus phylogenetic dissociation of the gene groups (as measure by the D value). The D value measures the phylogenetic clustering signal indicating a measure of a phylogenetic signal of a binary trait. The D value for each gene was calculated using coinfinder v1.1.0 (Whelan et al., 2020). For highly associated genetic features that fell in hypothetical genes, extra criteria were included such as the quality of predicted protein model and quality of Protein Data Bank (PDB) match (Berman et al., 2000) or blastp match, to assist elucidation of function.

3.2.7. Identification of hypothetical genes

Any protein annotation with no putative function was termed 'hypothetical'. The nucleotide sequence of each hypothetical (n=49) was extracted and a blastx carried out against the non-redundant (nr) protein sequence database to identify similarities to annotated proteins of known function (Altschul et al., 1990). For proteins (n=17) that remained hypothetical, protein modelling was undertaken. Each primary amino acid sequence was submitted to the Alphafold 2.0 software package through the google-collab web interface (Jumper et al., 2021) and each predicted protein model was visualised through PyMol v2.1 (Janson et al., 2017). The confidence measure pLDDT, a per-residue confidence metric, was used to evaluate the quality of the generated predicted protein models. The predicted protein models were then used as input into the web-interface DALI server to identify similar structures from the Protein Data Bank (PDB) (Holm and Laakso, 2016).

3.2.8. Identifying pStv which carries adhesin Stv

To determine the context and copy number of the putative novel adhesin Stv within the *Shigella* genome (Genbank ID = OP066525), both *in silico* and *in vitro* investigations were conducted for the clinical isolate corresponding with ERR1364216 (Baker et al., 2018b) which had been previously sequenced and contained the adhesin (screened for via abricate

utilising a custom database) (Seemann). To determine whether copy number differed for Stv and the chromosome, paired end reads were mapped back onto the assembly via Burrows-Wheeler Aligner (BWA) v0.5.9-r16 (Li and Durbin, 2009). The mapping files were filtered and sorted using samtools v1.13 (Li et al., 2009a). Coverage was calculated utilising samtools depth v1.13 and average calculated for both the chromosomal contigs and the small adhesin containing contig. Full reconstruction of the pStv sequence (Genbank ID = OP113953) was made using BLAST and ACT comparisons of the Stv-containing contiguous sequences in multiple isolates and the sequence confirmed through PCR analysis of DNA extracted from isolate ERR1364216 (Supplementary Table 1).

3.2.9. Protein tree construction

To explore the natural distribution of the novel adhesin Stv in other bacterial species, the amino acid sequence of Stv was utilised to carry out a blastp search against the clustered nr database. Sequences belonging to the cluster identified in this search (Steinegger and Söding, 2017) were extracted. The extracted protein sequences and the protein sequence of Stv were aligned via muscle v5.1 (Edgar, 2004) and RAxML-ng v8.2.9 (Kozlov et al., 2019) used to construct a phylogeny based on the protein alignment with 100 bootstraps.

3.2.10. Characterisation of key SNPs within the T3SS

To explore the SNPs within parts of the T3SS, each primary amino acid sequence of the wildtype T3SS protein and the SNP version of the T3SS protein were submitted to the Alphafold 2.0 software package through the google-collab web interface (Jumper et al., 2021) and each predicted protein model was visualised through PyMol v2.1 (Janson et al., 2017). The confidence measure pLDDT, a per-residue confidence metric, was used to evaluate the quality of the generated predicted protein models. The predicted protein models were then used as input into the web-interface of DeepDDG to calculate a $\Delta\Delta G$

score which was used to characterise the impact of the SNP on protein function and stability (Cao et al., 2019).

3.3. Results and Discussion

3.3.1. The Murray isolates are a valid continuum of Shigella *over time*

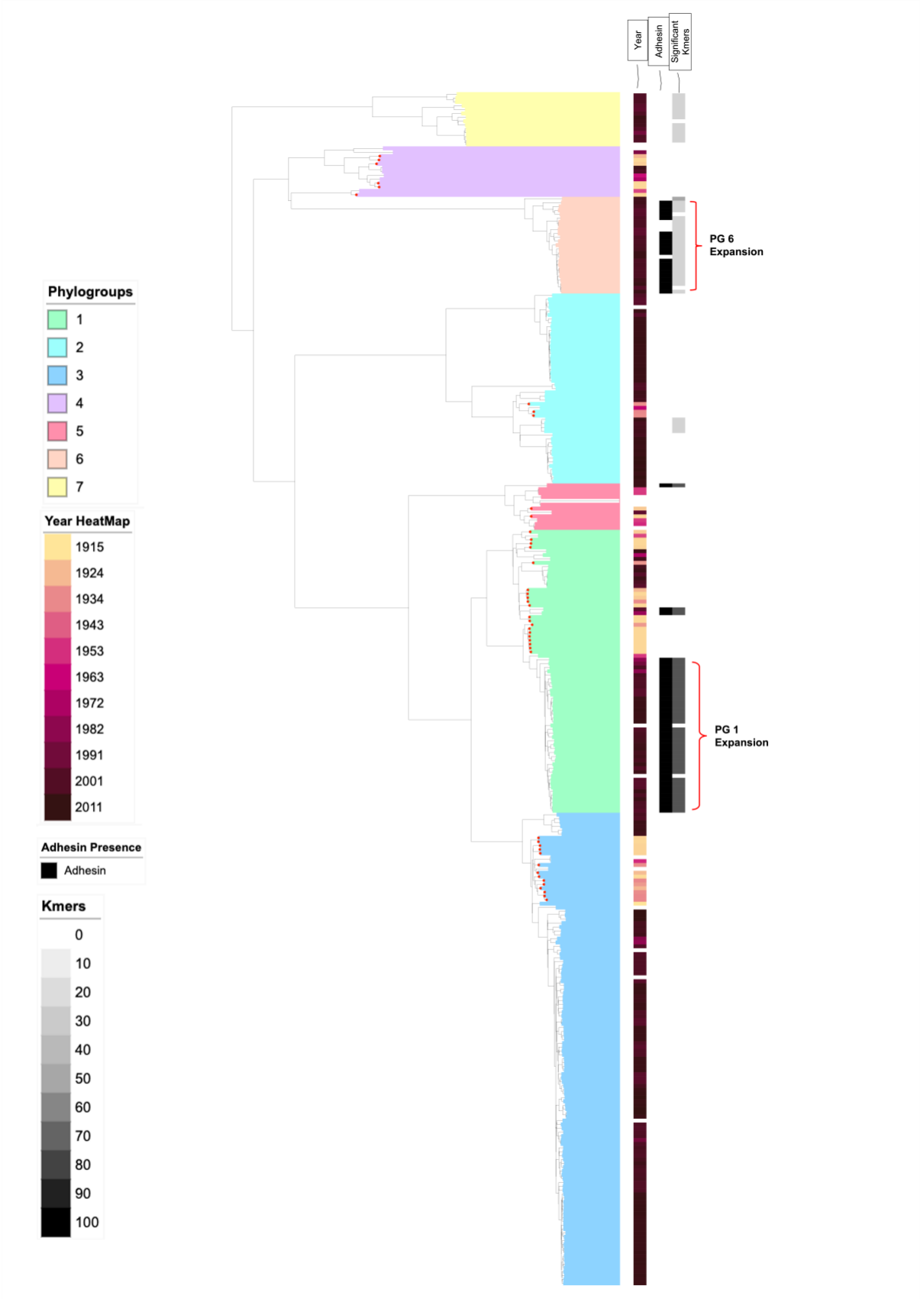


Figure 9: Contextualisation of historical isolates within the modern *S. flexneri* population structure and relationship with putative adhesin Stv.

The tree is a midpoint rooted maximum likelihood phylogeny for 309 *S. flexneri* isolates. Red circles overlaying tree tips indicate isolates from the historical Murray Collection. Phylogenetic groups (PGs) are shown shaded on the phylogeny and coloured according to the inlaid key. The year of recorded isolation is depicted as a heatmap in the colour strip closest to the tree and the presence/absence of adhesin Stv (according to abricate analysis) and number of significant kmers (from fsm-lite analysis) as a heatmap in the subsequent rightmost strips. Two phylogroup expansions of isolates containing Stv are indicated to facilitate discussion.

To compare the population structure of the pre-antibiotic era *Shigella* with modern *Shigella*, Murray collection *S. flexneri* isolates (n=45) were constructed into a phylogeny alongside a modern collection S*. flexneri* (n=262) representing all seven phylogroups (PGs) (Connor et al., 2015). Phylogenetic reconstruction using 84,391 SNPs identified in core regions revealed that Murray collection isolates belonged to PGs 1 through 5 but were absent in PGs 6 and 7 (Figure 9). Identifying the Murray *S. flexneri* isolates in a wide range of PGs suggests that *S. flexneri* PGs exhibit both co-existence and clonal replacement (i.e. the emergence of new *S. flexneri* PGs does not displace older PGs) consistent with previous studies (Connor et al., 2015). Indeed, most PGs contained a wide temporal range of isolates, indicating long term co-existence. These phylogenetic results suggest that historical and modern *S. flexneri* have similar chromosomes, further highlighting the likely importance of the accessory genome.

3.3.2. AMR and virulence determinants increase in time in S. flexneri

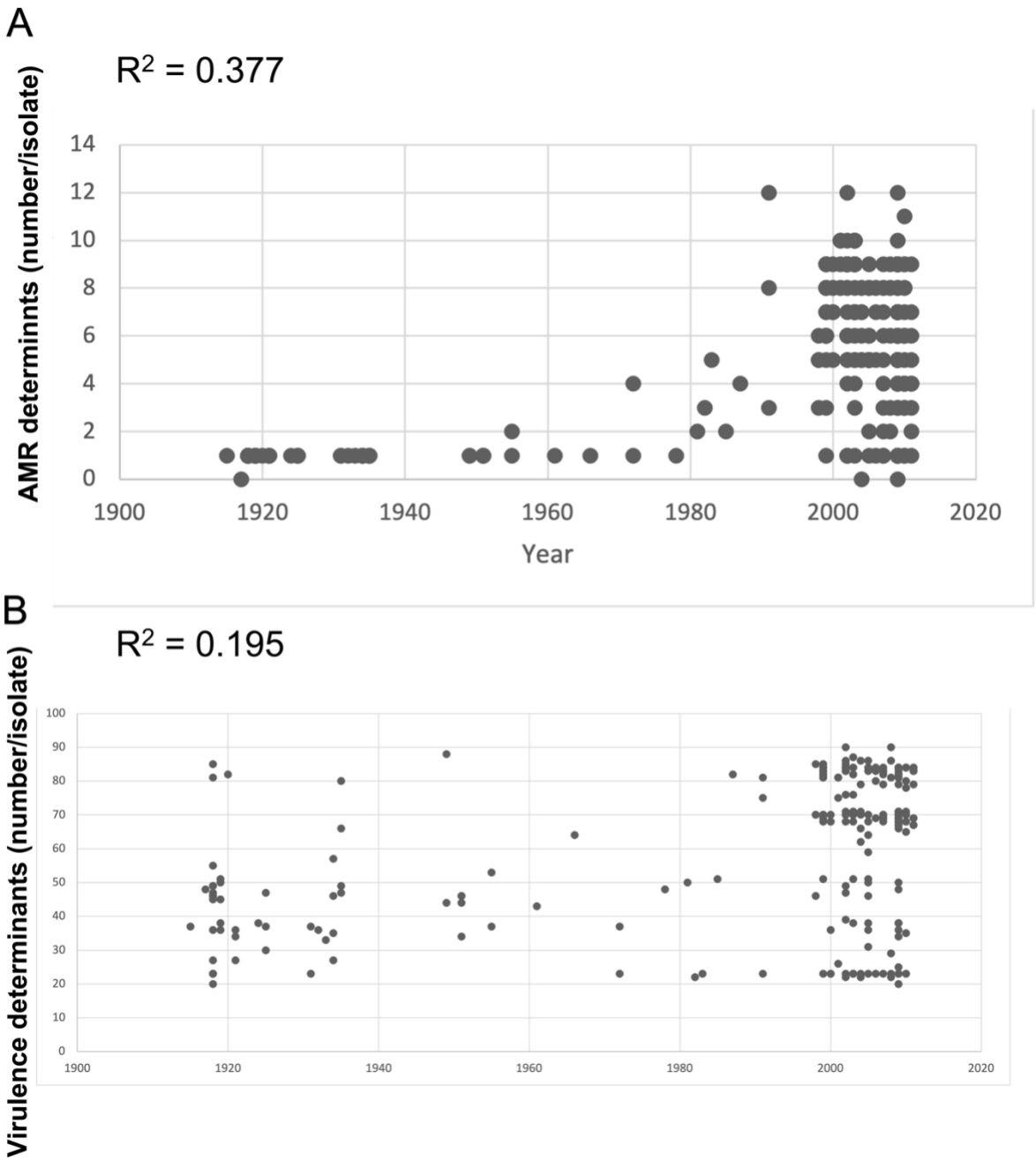


Figure 10: Temporal increase of AMR and virulence in *S. flexneri*.

A The number of AMR determinants per isolate is shown over time, with a linear trend line shown in blue. R^2 value noted at the top left. **B** The number of virulence factor (VF) determinants per isolate over time The linear trend line is shown in blue. R^2 value noted at the top left.

To investigate the evolution and acquisition of AMR in a time frame spanning the pre-antibiotic era, genotypic AMR profiles were determined for the Murray *S. flexneri* isolates revealing that all isolates contained a single AMR determinant (either *blaEC-8* (n=43) or *blaEC-15* (n=2), Supplementary Table 3).

To explore AMR trends over the century-long time span, genotypic prediction was conducted on the collection spanning 1915-2011 which revealed an accumulation of AMR over time, with some variation (Supplementary Table 3, Figure 10A). Statistical testing confirmed a significant difference in the mean number of AMR determinants per isolate between the Murray (n=45, $\bar{x}$=1) and modern isolates (n=246, $\bar{x}$=6.35, t_{291} = 11.6, *P* <0.001). This increase in AMR determinants over time is consistent with previous descriptions of *S. dysenteriae* Type 1 over similar time frames and broader trends in bacterial populations (Njamkepo et al., 2016b). This highlights that AMR is an appropriate marker for an accessory genome function that is increasing in *S. flexneri* within the time frame of this dataset.

As there was an increase in virulence determinants in the NCTC1 study, I also explored an increase in virulence genes over time. Similar to the *blaEC* genes, there were a selection of virulence factors that were encoded in >90% of the Murray *S. flexneri* isolates which had persisted in the modern isolates encoding for varying functions relating to host cell invasion, iron uptake, serotype conversion and porins (Table 3, Supplementary Table 3). All of these virulence factors have functions intuitively beneficial to successful *S. flexneri* infection.

Overall, there was a trend of an increase in virulence genes over time that was similar to AMR though with higher variation (Figure 10B). The differences between the number of virulence factors per isolate for the Murray *S. flexneri* (n=45, $\bar{x}$=42.4) and modern isolates (n=246, $\bar{x}$=65.8) was statistically significant (t_{291}, = 7.15, *P* = <0.001) which supports a signal of an increase in virulence factors over time. Although a general trend towards increasing virulence is less well described in bacteria, there are some precedents (such as the acquisition of serum resistance in invasive *Salmonella* Typhimurim (Hammarlöf et al., 2018), and increased phenotypic virulence in *Klebsiella* (Wand et al., 2015a)), and it is intuitive that

greater increase in virulence would contribute to continued public health visibility of a pathogen, despite vastly improving healthcare over the last century.

The increase in AMR and virulence over time in *Shigella* highlights their key roles in the success and visibility of *Shigella* as a pathogen. This result also indicates that AMR and virulence can be used as positive control functions for agnostically investigating functionally relevant genes that contribute to the long-term success of *Shigella*.

Table 3: Virulence factors found among historical *S. flexneri* isolates.

Table shows genes present in >90% of the Murray Collection *S. flexneri* isolates. Further detail in Supplementary Table 3.

Virulence genes found in pre-antibiotic era (Murray) *S. flexneri*	**Function**	**Relevant reference**
csgB, csgD, csgF	Host cell invasion	(Sakellaris et al., 2000, Hawkey et al., 2020)
espL4, espX4, espX5	Regulation of cell cycle	(Beltrametti et al., 1999)
fepA, fepB, fepC, fepD, fepG	Iron Uptake	(Schmitt and Payne, 1988, Fisher et al., 2009)
fes	Iron uptake	(Wei and Murphy, 2016)
fimA-I	Host cell invasion	(Klemm and Schembri, 2000, Chanin et al., 2019b)
gtrA, gtrB	Serotype Conversion	(Korres et al., 2005, Liu et al., 2008b)
iucA-D, iutA	Iron uptake	(Carbonetti and Williams, 1984)
ompA	Porin/essential for conjugation/bacteriophage receptor	(Pore et al., 2011)

3.3.3. *tGWAS reveals multiple genetic features positively associated with time in* S. flexneri

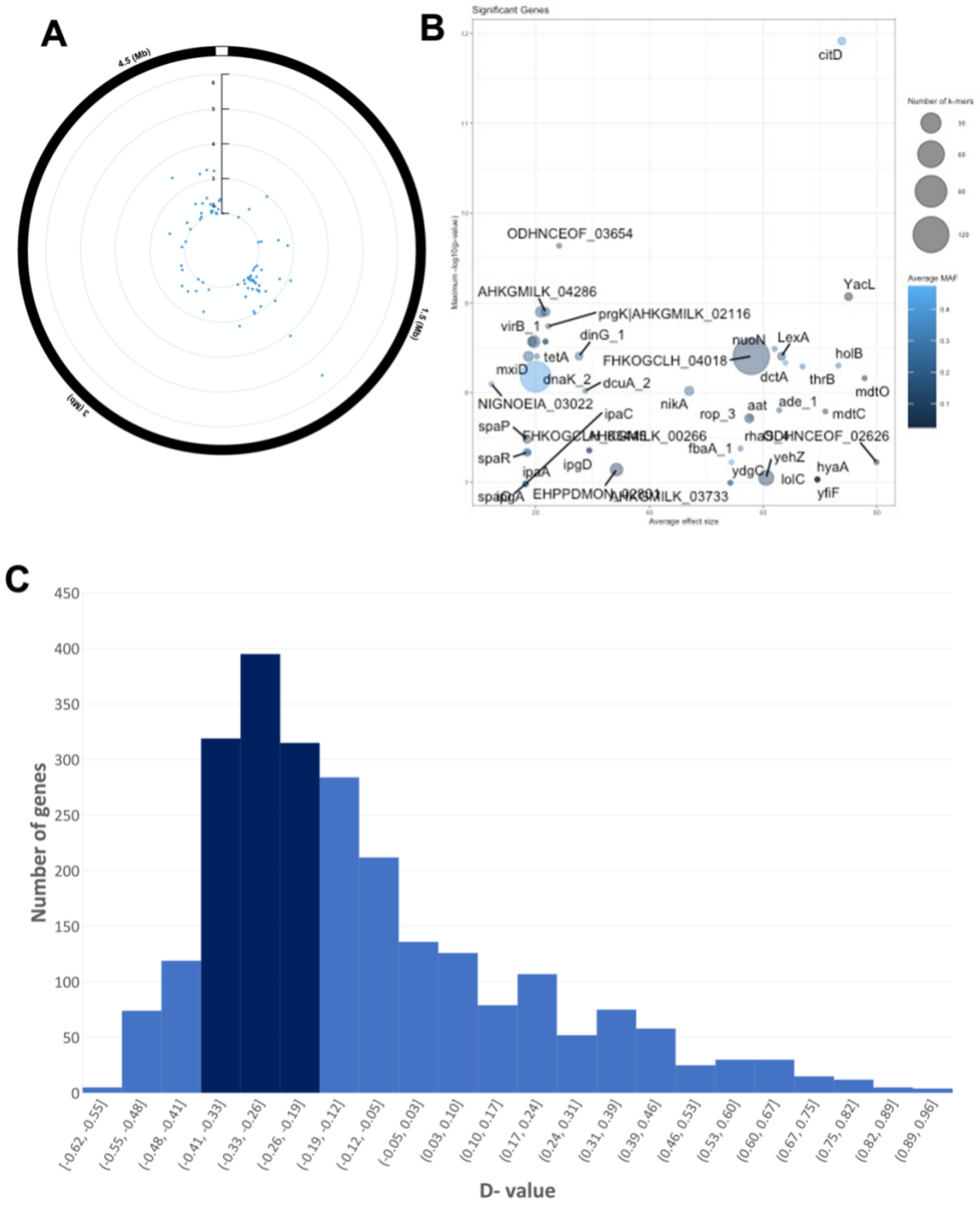


Figure 11: Genetic feature association in tGWAS by genetic feature type including SNPS (A), kmers (B), and COGS (C).

A SNPs: The circular manhattan plot shows the genomic position (clockwise) and negative logarithm of the p-value (radial axis) for 93 significantly positively associated SNPs **B kmers:** Bubble plot showing the number of kmers (bubble size) by gene (text labels in graph field), effect size (x-axis), and statistical support (y-axis as negative logarithm of the p-value). **C COGs:** Phylogenetic dissociation distributions for all COGs identified as positively and significantly associated with time in tGWAS (light blue) and the 25 COGS with the strongest statistical support (the dark blue bins).

Conducting bacterial GWAS with time as a continuous variable (coined herein as tGWAS) identified numerous significant SNPs (n=93, minlog10(p)>2), kmers (n= 306, p-value<1.04E-07) and COGs (n=359, lrt p-value<0.05) (Figure 11, Supplementary Table 4) which were explored further using a custom decision tree (Supplementary Figure 1).

In evaluating the likely impact of the SNPs, I focused on those within genes (totalling 93 SNPs within 48 genes). Of these, 83% (n=77/93) were located on the chromosome and the remaining 17% (n=16/93) were located on the virulence plasmid (VP). This overrepresentation (with respect to size i.e., chromosome is 4607202 bp and virulence plasmid (221618 bp) of SNPs associated with time on the VP relative to the chromosome may indicate increased importance of VP evolution in *Shigella*. Among the associated SNPs in genes, 38% (n= 35/93) encoded a missense mutation or stop codon which may potentially impact protein structure so were selected for further study (Table 2, Supplementary Table 3). Regarding the kmer and COG analyses, sequence kmers of various lengths (10 – 100 bp) were identified in 47 genes; 85% (n=40/47) which encoded proteins of known function and 7 proteins of unknown function (Table 4, Supplementary Table 4). Notably, two T3SS genes (ipaC and ipgD) were identified among both the 48 genes identified with SNPs and 47 genes containing significant kmers (Table 4). tGWAS based on COGs returned 359 significant groups based on one method of COG clustering (roary) which only reduced to 228 when cross referenced with another clustering method (panaroo) (Supplementary Table 3) (Tonkin-Hill et al., 2020). The vast quantity of COGs (>200) posed an issue for investigation as there was no high-throughput method for the further in-depth analysis of each COG. Pragmatically, the 25 COGs with the highest statistical significance of association (as measured by lrt p-value) from the primary analysis that were also present after cross-referencing were taken forward for further analysis (Table 4, Supplementary Table 4), only 20% (n=5/25) of which were genes

with known functions. One gene*, rop_3*, was found through tGWAS based on both kmers and COGs (Table 4). The plethora of genetic factors significantly associated with *Shigella* in time offers a unique perspective on the wide variety of factors that may be contributing to the success of *S. flexneri* as a pathogen.

Table 4: Gene-associated genetic features associated with time in *S. flexneri*

Genes and whether they were identified in SNP, kmer, or COG (25 most significantly associated) tGWAS analysis, as well as whether they form part of the T3SS and their known function and related references are shown. ND – not described or investigated as part of this study. Additional details found in footnote 2 and Supplementary Table 4.

2

Gene	SNP	Kmer	COG	T3SS	Function	Reference
imp	√				Hypothetical	ND
insB	√				Insertion element IS1 protein insB	(Ohtsubo et al., 1981)
yncB	√				Curcumin-converting enzyme	(Hassaninasab et al., 2011)
tehB	√				Tellurite resistance protein	(Turner et al., 1995)
aldA	√				Aerobic dehydrogenase	(Vergara-Irigaray et al., 2014)
yeeE	√				Permease?	ND
hmpA	√				Nitrosative stress	(Eriksson et al., 2003, Lucchini et al., 2005)
yihQ	√				Alpha-glucosidase	
sgaU	√				Sugar Metabolic Pathway	(Han and Lee, 2006)
ipaC	√	√		√	T3SS	(Terry et al., 2008)
ipgD	√	√		√	T3SS	(Niebuhr et al., 2000)
yiiX	√				Peptidase?	(Yang et al., 2005)
SF1013	√				Pyrimidine utilisation?	ND
SF1015	√				Pyrimidine utilisation?	ND
SF1742	√				Hypothetical	ND
SF1803a	√				Hypothetical	ND
SF4472	√				Hypothetical	ND
SF1928	√				IS3 Family Transposase	(Hawkey et al., 2020)
SF2042c	√				Hypothetical	ND
SF2581	√				Hydrolase?	ND
SF2866	√				Integrase?	ND
SF3181	√				Chaperone?	ND
SF3255	√				Transposase?	ND
SF4259	√				IS4 family transposase	(Yang et al., 2005)
CP0007	√				IS3 family transposase	(Hawkey et al., 2020)

[2] The known gene names are taken from the annotated *S. flexneri* 2a strain 301 reference genome and unknown genes entitled SF or CP were also named from the reference genome and its virulence plasmid. Other unknown gene names were assigned based on the first genome they were observed in from the isolate collection.

CP0017	√				IS4 family transposase	(Yang et al., 2005)
CP0019	√				IS6 family IS element TnpB	(Siguier et al., 2014)
CP0217	√				CopG family protein	(Solar et al., 2002)
dinG_1		√			DNA helicase	ND
ODHNCEOF_02626		√			Hypothetical	ND
mdtC		√			Multidrug resistance	(Horiyama and Nishino, 2014a, Pletzer and Weingart, 2014)
ODHNCEOF_03654		√			Hypothetical	ND
citD		√			Citrate metabolic pathway	ND
mdtO		√			Multidrug resistance	(Horiyama and Nishino, 2014a, Pletzer and Weingart, 2014)
icsB		√		√	T3SS	(Mattock and Blocker, 2017)
AHKGMILK_04072		√			Hypothetical	ND
spaO		√		√	T3SS	(Bajunaid et al., 2020)
mxiD		√		√	T3SS	(Notti and Stebbins, 2016, Martinez-Argudo and Blocker, 2010)
spaP		√		√	T3SS	(Bajunaid et al., 2020)
virB_1		√		√	T3SS	(Bajunaid et al., 2020)
AHKGMILK_04286		√			Hypothetical	ND
ipaA		√		√	T3SS	(Mattock and Blocker, 2017)
prgK|AHKGMILK_02116		√		√	T3SS	(Muthuramalingam et al., 2021)
AHKGMILK_00268		√			Phage related gene	ND
mxiC		√		√	T3SS	(Notti and Stebbins, 2016, Martinez-Argudo and Blocker, 2010)
spaR		√		√	T3SS	(Bajunaid et al., 2020)
rhaS_4		√			Transcriptional activator	(Bhende and Egan, 2000)
aat		√			Transferase	ND
fbaA_1		√			Adolase	(Bardey et al., 2005)
nuoN		√				
AHKGMILK_01446		√			General stress protein	ND
AHKGMILK_00266		√			Phage related gene	ND

AHKGMILK_03733		√			Type II T/AT system	ND
ydgC		√			Alginate biosynthesis	ND
AHKGMILK_03728		√			Transcriptional regulator	ND
ade_1		√			Hypothetical	ND
ipgA		√		√	T3SS	(Mattock and Blocker, 2017)
FHKOGCLH_04018		√			Now characterised as an Adhesin	ND
dnaK_2		√			Chaperone	ND
tetA		√			Tetracycline resistance protein	(Møller et al., 2016)
rop_3		√	√		Plasmid regulation	(Cesareni et al., 1982)
dctA		√			Transport protein?	ND
FHKOGCLH_04017		√			Hypothetical	ND
dcuA_2		√			Hypothetical	ND
FHKOGCLH_03445		√			Hypothetical	ND
yehZ		√			Permease?	ND
EHPPDMON_02801		√			Hypothetical	ND
lolC		√			Release of lipoproteins	(Tang et al., 2021)
yfiF		√			Transcriptional regulator	ND
hyaA		√			Hydrogenase	(Menon et al., 1990)
NIGNOEIA_03022		√			DNA Helicase	ND
holB		√			DNA polymerase subunit	ND
thrB		√			Homoserine Kinase	ND
NIGNOEIA_02718			√		YgcB	ND
focC			√		Fimbriae formation	(Riegman et al., 1990)
IAAIPMMA_04017			√		Hypothetical	ND
EHPPDMON_02080			√		YeaG - kinase	ND
NIGNOEIA_02490			√		Phage related genes	ND
entS			√		Enterobactin exporter	(Furrer et al., 2002)
JHKOGANH_03285			√		Mercury resistance	ND
IKOJHFDB_04319			√		Phage related genes	ND
IKOJHFDB_04422			√		Peptidase?	ND
AOKKNEDP_03748			√		IS4 Family transposase	(Hawkey et al., 2020)
BMAIACIG_03405			√		Hypothetical	ND
AHKGMILK_04030			√		Hypothetical	ND
AHKGMILK_04155			√		Phage related gene	ND
NLBACPAE_04095			√		Hypothetical	ND

IAAIPMMA_03532			√		Transcriptional regulator	ND
PNNKJLLM_03516			√		disT?	ND
KGMACLPL_03970			√		Replication regulator	ND
KGMACLPL_03971			√		Plasmid Recombination	ND
nanM			√		Sialic acid metabolic pathway	(Steenbergen et al., 2009)
nanS			√		Sialic acid metabolic pathway	(Steenbergen et al., 2009)
chbC			√		Chitobiose metabolic pathway	(Verma and Mahadevan, 2012)
AHKGMILK_03727			√		Hypothetical	ND
AHKGMILK_03728			√		Transcriptional regulator	ND
AHKGMILK_03730			√		Hypothetical	ND[3]

3.3.4. tGWAS confirms antimicrobial resistance as increasing and evolving in time

Having gathered all potential genetic features associated with time by tGWAS, I examined whether AMR determinants were recovered (Table 4). AMR was clearly recovered by tGWAS, acting as an effective positive control for the identification of functional accessory genome determinants that contribute to the pathogen success over time. Significant kmers (n=3) were found within several AMR determinants, namely *mdtC*, *mdtO* and *tetA* (Table 4). mdtC is a resistance protein forming part of the multidrug efflux complex MdtABC which provides resistance to aminocoumarins antimicrobials, is involved in enterobactin export, and is involved in transport of flavonoids, fusidic acid, josamycin, bile salts and silver nitrate (Horiyama and Nishino, 2014b, Pletzer and Weingart, 2014). Similarly, *mdtO* encodes for a resistance protein conferring resistance to puromycin, acriflavine and tetraphenylarsonium chloride and *tetA* is responsible for tetracycline resistance (Møller et al., 2016). Genes

[3] The known gene names are taken from the annotated *S. flexneri* 2a strain 301 reference genome and unknown genes entitled SF or CP were also named from the reference genome and its virulence plasmid. Other unknown gene names were assigned based on the first genome they were observed in from the isolate collection.

conferring resistance to antimicrobials and bile salts have clear beneficial roles for a human host-adapted intestinal pathogen where antimicrobials are used in treatment and bile salts are a key part of the intestinal environment. AMR related genes also were recovered in the COG analysis, including multiple *mdt* and *tet* genes, but these were not present within the 25 COGs with the lowest lrt p-value (<2.62E-02)

3.3.5. S. flexneri *virulence determinants, including the T3SS, are changing in time*

Virulence factor genes were also expected to be associated with success over time, and the T3SS is a key virulence mechanism that encodes the machinery required for *Shigella* to invade the colonic epithelium (Sansonetti et al., 1982). Ongoing evolution of the T3SS was indicated by significant SNPs and kmers being found during tGWAS (Table 4). Specifically, there were singular missense SNPs within *ipaC* and *ipgD* (Supplementary Table 4) encoding invasion antigen C that initiates actin mobilisation and the early steps of membrane ruffling, and a T3SS effector protein thought to modulate host cell response respectively (Terry et al., 2008, Niebuhr et al., 2000). Further support for the T3SS role as a key virulence mechanism over time was shown by 12 of the 47 genes in which significant kmers were found were associated with T3SS machinery, specifically in *icsB*, *spaO/P/R/*, *mxiC/D*, *virB*, *ipaA/C*, *ipgA/D*, and *prgK* (Table 4, Supplementary Table 4). Many different parts of the T3SS machinery were identified, including regulators (*virB*, *mxiC*), assembly machinery (*spa* genes), effector proteins and their chaperones (*ipaA/C*, *icsB*, *ipgA*). Having the T3SS genes observed in multiple tGWAS feature types highlights the essential and changing nature of the T3SS to *Shigella* success over time. tGWAS hits relating to a variety of other virulence, invasion, and metabolic processes were also detected. These included: singular SNPs in the nitrous oxide and tellurite resistance genes *hmpA* and *tehB*; a SNP in the virulence plasmid gene CP0217 and kmers and COGs in

rop_3, both being genes that likely contribute to efficient plasmid replication and retention; the *nanM* and *nanS* genes involved in sialic acid binding and metabolism appearing in the 25 most-significant COGs; and high support for kmers in the metabolic *citD* gene which might help *S. flexneri* exploit resources (Table 4, Figure 11). For survival and successful infection, *S. flexneri* must counteract the high concentrations of toxic chemicals and maintain key plasmids for invasion. Thus, the known gene products identified have intuitively helpful functions for the virulence of *Shigella* over time, again highlighting the validity of tGWAS.

3.3.6. Insertion elements revealed as a key contributor to the success of S. flexneri *over time*

IS expansion is thought to be an early step in the genome reduction process adopted by bacteria that have recently adopted host-restricted lifestyles (Siguier et al., 2014). *Shigella* was the first example of IS expansion within bacterial genomes and since then many IS families have been documented throughout bacterial genomes (Nyman et al., 1981). Individual significant SNPs fell within genes that encoded proteins involved with insertion elements (*insB*, SF1928, SF4259) (Table 4, Supplementary Table 4). The insertion element 1 protein InsB, SF1928 and SF4259 (IS transposases for IS3 and IS4) all contribute to strain evolution, variation, and functional gene loss and streamlining (Hawkey et al., 2020). Remarkably, there were also singular SNPs within transposases on the virulence plasmid, CP007 and CP0017 (related to IS3 and IS4 respectively) (Table 4, Supplementary Table 4). The prominent presence of SNPs occurring within the IS related genes (n=9 10% of total SNPs) further supports their inferred importance for *S. flexneri* over time. To further support their importance for *Shigella's* evolution AOKKNEDP_03748 was observed as a significant COG which encodes for an IS4 family transposase. IS elements within a host adapted organism such

as *Shigella* are intuitively beneficial by contributing to functional gene loss and genome streamlining facilitating host adaptation.

3.3.7. *Identifying Stv: a novel adhesin among genes of unknown function*

Not all significant tGWAS hits were in characterised genes. Specifically, 6% (n=6/93) of the associated SNPs, 15% (n=7/47) of the kmers, and 80% (n=20/25) of the top COGs were in genes of unknown function, which may represent other key genes in the long-term success of *Shigella*, given that they appeared alongside AMR and virulence.

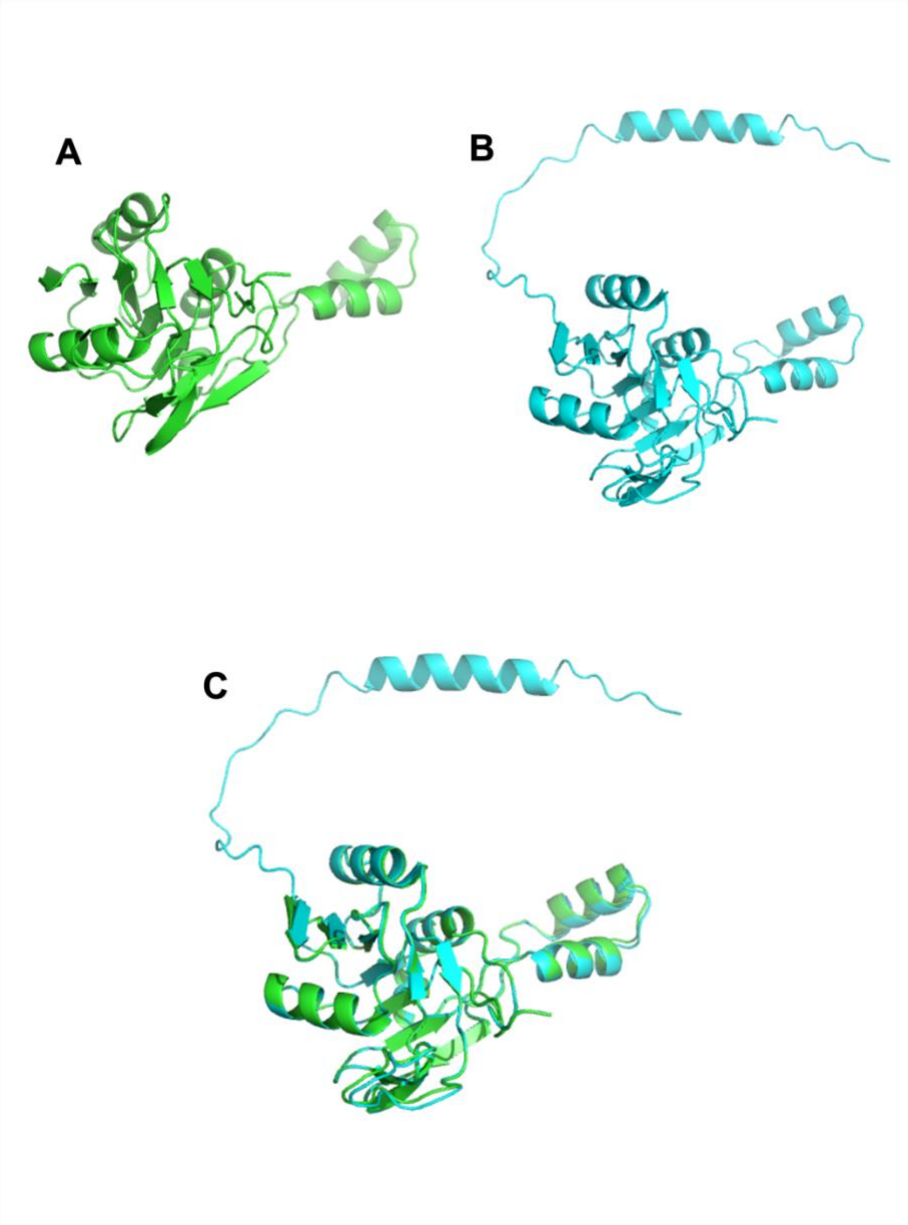


Figure 12: Modelling the undescribed adhesin Stv that is associated with *Shigella* in time

A Predicted Alphafold model for FHKOGCLH_04018 (green) visualised in Pymol. **B** Predicted Alphafold model for *E. coli* adhesin (WP_205849698.1) (cyan) visualised in Pymol. **C** Alignment of modelled FHKOGCLH_04018 (green) and *E. coli* adhesin (cyan) (WP_205849698.1).

Numerous significant kmers (n=123, of length 12 – 100 bp) were located within hypothetical protein FHKOGCLH_04018 with a large effect size (Figure 11B). BLASTp searches found a 99% amino acid sequence identity match between FHKOGCLH_04018 and a putative *E. coli* adhesin protein WP_205849698.1 (no current publications associated). Both protein sequences were then modelled using Alphafold (Jumper et al., 2021) (Figure 12). FHKOGCLH_04018 folded into a seven stranded β-sheet, which was predominantly anti-parallel, with four α-helices on one side and a single α-helix on the other (Figure 12A). The resulting model of WP_205849698.1 showed clear structural similarity with FHKOGCLH_04018, with the majority of the fold aligning well, with the exception of the C-terminal helix that was folded in WP_205849698.1 but modelled with only very low confidence as random coil in FHKOGCLH_04018 (Figure 12B). An all-atom root-mean-square deviation of atomic positions (RMSD), calculated by running a structural superposition of the two models, produced a score of 1.498 Å (<2 indicates very close alignment, and this value was calculated based on the alignment of 3264 atoms between the two models). Due to this stark similarity, FHKOGCLH_04018 was also assigned putative function as a novel *S. flexneri* adhesin, herein called adhesin Stv.

Subsequent analyses revealed that Stv was carried on a small plasmid and was widely distributed in bacterial populations. In draft *S. flexneri* genomes, Stv occurred on small contiguous sequences, ~2.5 kb in length, with two other genes. The putative plasmid was reconstructed using multiple draft genome comparisons and the 2689 bp element was confirmed using bidirectional coverage PCR and is herein termed pStv (Supplementary Figure 2). Relative mapping depth of the chromosome and pStv in the Stv containing clinical isolate of *S. sonnei* (methods) was calculated. This revealed that pStv occurred at an approximate

ratio of 8.6:1 relative to the genome. The higher coverage for pStv suggests that the novel adhesin is found on a small multi-copy plasmid. pStv contains two other genes including another uncharacterised gene, *FHKOGCLH_04017*, and *rop_3*, the second most significant COG in the tGWAS analysis. As per the methodology set out for hypothetical genes, efforts were made to identify the function of *FHKOGCLH_04017,* however, even after a good quality protein predicted model was generated no quality match was found on the PDB or via blastp searches and so remains uncharacterised.

3.3.8. The significance of Stv in Shigella *and in other bacteria*

To infer the potential relationship of Stv with *S. flexneri* population dynamics, the distribution of Stv was explored within the *S. flexneri* phylogeny. Specifically, from the phylogeny it can be observed that Stv is polyphyletic (Figure 9), highlighting that Stv has been introduced convergently. This polyphyletic nature was demonstrated through the presence of Stv largely in Phylogroups 1 and 6 (Figure 9), the latter of which contained the earliest Stv-containing (from 1978). Variations in the number of Stv kmers suggested that there may be variants of Stv in other PGs. And notably, Stv appeared to have been associated with PG expansions, where acquisition was followed by clonal expansion, particularly for PG1 (Figure 9). Collectively, these findings support that Stv may have an important role in *S. flexneri* poluation dynamics.

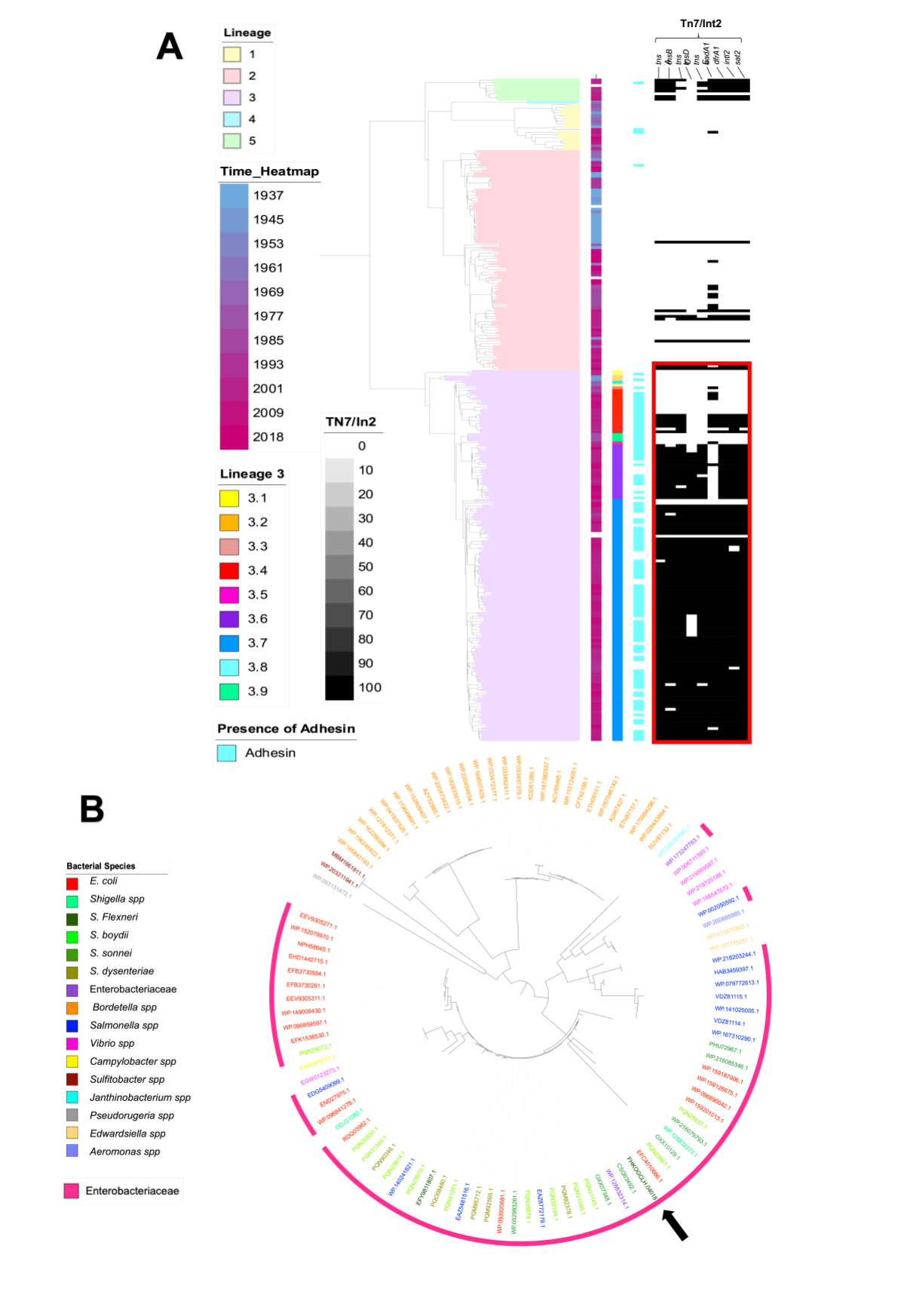


Figure 13: The natural distribution of Stv in other bacterial species.

A Mid-point rooted phylogeny of 305 *S. sonnei* isolates. Coloured clades indicate the *S. sonnei* lineages. Coloured strips depicting Lineage 3 genotypes, presence of Stv and presence of the AMR genes from the Tn7/Int2 cassette **B** Phylogeny of aligned protein sequences of the most closely related clustered adhesin sequences. Coloured tips represent the different bacterial species and Enterobacteriaceae species are highlighted via the coloured strip. The black arrow depicts the presence of the *S. flexneri* adhesin Stv.

To externally validate the potential contribution of Stv beyond *S. flexneri*, I related the presence of Stv to the next most burdensome *Shigella* species, *S. sonnei*. *S. sonnei* has comparatively well described natural history and known drivers of clonal expansions including AMR and colicin acquisition (Holt et al., 2012b, Holt et al., 2013, Hawkey et al., 2021) Stv was found sporadically in Lineages 1, 2, and 5, but was acquired and stabilised early in the emergence of Lineage 3 (Figure 13A). Lineage 3 of *S. sonnei* is the globally dominant multidrug resistant (MDR) lineage whose success is thought to be attributable to the acquisition of key AMR determinants, particularly the Tn7/Int2 cassette (Figure 13A) (Holt et al., 2012b). Our results indicate that Stv predates the acquisition of the Tn7/Int2 MDR determinant. Specifically, the first observation of the adhesin within Lineage 3 was in 1943 while the first observation of any part of the Tn7/Int2 cassette was in 1990 suggesting that this may have provided the opportunity for Stv-positive *S. sonnei* to further mobilizable elements such as AMR and colicins, ultimately combining to drive the success of this global MDR lineage. This represents an exciting insight into the potential contribution of this putative novel adhesin in both *S. flexneri* and *S. sonnei*.

Owing to its importance in shigellae, I searched for close relatives of Stv in the wider context. I found clustered relatives of Stv in the nr database across multiple other bacterial species (Figure 13B). Including other *Shigella* species, related *Enterobacteriaceae* (e.g. *E. coli*, *Salmonella*), and more distantly related bacteria (e.g. *Bordetella spp*) indicating widespread conservation, and the possible importance of Stv in shaping the population structures of other important AMR pathogen groups.

Despite concerted effort, I were unable to find any functional experiments for the putative *E. coli* relative of Stv in the literature. Our *in-silico* investigation highlights the importance of the previously uncharacterised Stv adhesin for the success of *S. flexneri* in time and its potential importance in other species. Functional work to confirm the predictive adhesive function of the protein is ongoing but beyond the scope of this study. Once confirmed, Stv would represent a potential drug target as anti-adhesion therapies have the potential to be effective in the prevention and treatment of bacterial infections (Asadi et al., 2019). Our encouraging results with Stv indicate the likelihood of success that functional microbiologists may have in exploring the other genetic features I have identified for tGWAS for *Shigella* (Supplementary Table 4).

3.3.9. The challenges in characterising the SNPs within the T3SS

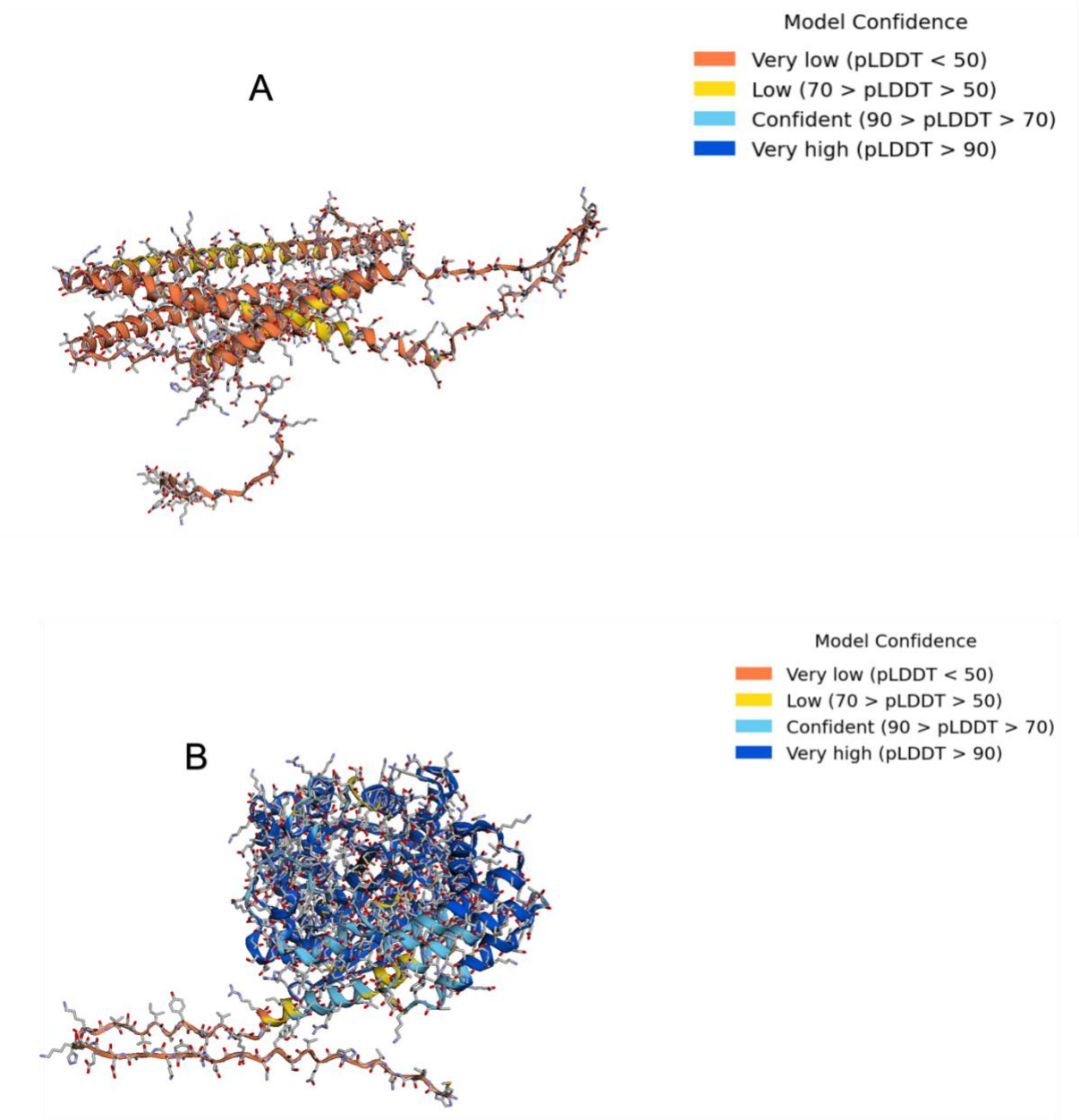


Figure 14: Protein modelling of key T3SS proteins.

A – Generated Alphafold model for ipaC. **B** – Generated Alphafold model for ipgD. Keys on the right hand side depict confidence measures of the predicted models.

Due to the importance and vital nature of the T3SS for successful Shigella infection and virulence, it was critical to investigate the singular significant SNPs found within *ipaC* and *ipgD*. To fully evaluate the effect that the SNPs were potentially having on the structure and hence function of the T3SS proteins it was imperative to view the SNPs through visualisation as a protein model. Even though ipaC and ipgD are key parts of the well described T3SS, after thorough searches of the PDB it yielded no solved protein structures for either ipaC or

ipgD. Due to the lack of a confirmed protein structure, predicted models of both T3SS proteins were modelled via Alphafold (Jumper et al., 2021). The model of ipaC (Figure 14A) was plagued with very low to low pLDDT values across the entire model inferring low confidence in the model as a whole. Despite the low confidence model, continued investigation to evaluate the effect of the SNP was conducted. The SNP within ipaC was located at bp position 868 of a 1091 bp length ipaC protein. The SNP converted a cytosine to a thymine (C -> T) resulting in a missense mutation where the amino acid at position 284 was converted from a glutamic acid to aspartic acid (E -> D) (Figure 14A). Previous work defining the critical functional regions of ipaC denotes the N-terminal region 50-80 as critical for ipaC invasion, whilst the C-terminal from amino acid 300 onwards contains regions critical to invasion, formation, and virulence (Harrington et al., 2003, Terry et al., 2008). There is a lack of functional information regarding the region around amino acid residue 283 and so it is difficult to conclude whether the location is important for critical functions of ipaC. Research deleting amino acids from 260 onwards does result in changes to the effectiveness of ipaC, however, the changes would likely result due to deletions of the C-terminal and not necessarily to do with the residue around the 284 region (Picking et al., 2001). The position of the missense mutation is on an alpha helix (Figure XX) and evaluation via DeepDDG calculated a predicted stability change ($\Delta\Delta G^{stability}$) of -0.391 kcal/mol. A $\Delta\Delta G$ score of between 0.5 to -0.5 is deemed to have a negligible effect on protein function or stability.

With a low confidence model and the lack of information regarding the functional region around residue 284 coupled with the low $\Delta\Delta G$ score, I cannot conclude that this SNP within ipaC is having a significant positive impact on the function of the protein.

Predictive modelling of ipgD (Figure 14B) resulted in a better quality and higher confidence model. The singular significant SNP occurred at bp 982 out of a gene that is 1616 bp in length, resulting in a guanine mutating to an adenine (G -> A). This change resulted in a missense mutation where a serine at residue 328 was altered to an asparagine (S -> N). In *Shigella*, there has been little to no characterisation of any of the functional regions within this protein. However, there has been structural characterisation of an orthologue of ipgD, SopB in Salmonella (PDB=4DID). The structural characterisation of SopB reveals key functional regions as a binding domain (residues 117-168) and a phosphoinositide phosphatase domain (residues 357-561) (Burkinshaw et al., 2012). Although exact residues and structural differences may be present between SopB and ipgD the general location for residue 328 does not seem to be a key functional region. Further investigation of the impact on the single mutation via DeepDDG calculated a $\Delta\Delta$G score of -2.355 kcal/mol. A score of < -0.5 is deemed destabilising and typically is not advantageous for protein structures. Although there was a better predicted model for ipgD, the location and $\Delta\Delta$G score indicate that the SNP is once again having a largely neutral effect on the protein structure and hence function. Similarly, to ipaC, functional work would be needed to confirm in further detail the exact impact of the SNP. Overall, in silico investigation on the impacts of SNPs are difficult to pinpoint exact impacts. There are multiple challenges including the confidence and accuracy of predicted models, identifying impacts of single mutations and identifying key functional regions of proteins. In silico investigations may need to be combined with functional work to fully evaluate the impact of key SNPs.

3.4. Conclusion

Through the investigations utilising the Murray Collection in tandem with modern isolates,

the importance of historical isolates as an important and rich resource of information to provide insights into bacterial evolution over time has been highlighted. Our development of tGWAS has generated a novel validated methodology and workflow for investigating genetic factors contributing to pathogen success over time. For *Shigella*, this highlighted the roles of antimicrobial resistance, virulence and invasion, genetic plasticity, and host colonisation through the identification of a previously unreported adhesin, Stv, in long term pathogen success. This putative adhesin likely has relevance for other bacterial species and is an elegant demonstration of the potential insights to be gained using our novel tGWAS methodology which is broadly translatable to other bacterial species.

3.5. Next Steps

Due to the importance of MGE's and interbacterial competition mechanisms seen throughout both tGWAS analyses of *S. sonnei* and *S. flexneri*, it seemed prudent to investigate a key interbacterial advantage noted in *S. sonnei* Lineage 3. Within *S. sonnei* Lineage 3 an *E. coli* killing phenotype had been observed which would be intuitively beneficial to the success of *Shigella* species as *E. coli* is a common bacteria in the host microbiota. The factors behind this advantageous ability remained unknown and so GWAS approaches were utilised to identify these factors.

Chapter 4

Much of the content of this chapter is to be published in the research article "*Escherichia coli* killing by epidemiologically successful sublineages of *Shigella sonnei* is mediated by colicins" which is currently under review. Permission to include the publication in this PhD was obtained from all co-authors. Specifically, I acknowledge the contribution of authors below. Unless specified below all work was completed by myself.

P. Malaka De Silva	Laboratory work including BPER via cell sorter and plate reader. Extraction of supernatants for mass spectroscopy.
Lauriene Kuhn	Completion of Mass spectroscopy
Patryk Ngondo	Completion of Mass spectroscopy
Brian Ho	Guidance on the functionality of the T6SS
Francois-Xavier Weill	Gave us the nanopore assembly of CIP
Benoît S. Marteyn	Supervision and guidance of the mass spectroscopy carried out by his group.
Lorine Debande	Completion of Mass spectroscopy
Kate S Baker	Supervision
Claire Jenkins	Guidance and strain selection

4. *Escherichia coli* killing by epidemiologically successful sublineages of *Shigella sonnei* is mediated by colicins

4.1. Introduction

Shigella sp. cause >125 million cases of shigellosis that result in ~212,000 deaths annually and understanding the biological drivers of its success as a pathogen is critical to public health (Bardhan et al., 2010, Khalil et al., 2018). *S. sonnei* accounts for 24% of cases and contributes proportionately more to the disease burden in industrialising and high-income countries (HICs) where it is primarily transmitted through travel and among certain risk groups e.g. men who have sex with men (MSM) (Baker et al., 2015b, Ingle et al., 2019, Thompson et al., 2015). Recent studies of *S. sonnei* have revealed a concerning emergence of highly, and extensively drug resistant strains (Simms et al., 2015, Bardsley et al., 2020, Charles et al., 2022, Chung

The et al., 2016, Fischer et al., 2021, Locke et al., 2021, Zayet et al., 2021, Moreno-Mingorance et al., 2021), leading to it being designated a WHO priority organism for AMR and highlighting the need to continue study of this important pathogen (Tacconelli et al., 2018). Global subtyping systems can help us trace the shifting dynamics, but there is a need to leverage these genomic epidemiology studies to better understand the biology of the organism by identifying what drives certain sublineages to success (Hawkey et al., 2021).

An example relevant to this study is the existence of four co-circulating subclades of MSM-associated *S. sonnei* in the UK during 2008 to 2014, three of which endured in time and spread internationally, and one of which (Clade 3, which belongs to genotype 3.7.18 under the *S. sonnei* typing scheme described above) did not (Baker et al., 2018b). Being an enteric pathogen, *S. sonnei* is in constant competition with the other members of the gut microbiota which acts as the barrier to successful infection and invasion of the gut epithelial cells in the colon (Anderson et al., 2016). Therefore, mechanisms that are advantageous for interbacterial competition are beneficial for *S. sonnei*.

There are a plethora of interbacterial competitions mechanisms including Type VI secretion systems (T6SSs), one of which was recently described in *S. sonnei* and colicins (Sorbara and Pamer, 2019, Anderson et al., 2017b). T6SSs are specialised secretion systems found in Gram negative bacteria capable of delivering a wide variety of effectors (usually antibacterial proteins) directly into the target cells (Coulthurst, 2019, Zoued et al., 2014). T6SSs function via a contractile sheath that propels a needle-like structure containing the spike complex with the effector protein into the adjacent target cell in a contact-dependent manner (Cianfanelli et al., 2016). Contrastingly to T6SS, colicins are not contact-dependent as they are secreted into the surrounding environment and translocated through the outer membrane of target cells by either passive diffusion or active transport (Lazdunski et al., 1998). These small toxic

proteins are produced by many enteric bacteria and are usually encoded on a plasmid alongside the colicin lysis protein that is responsible for colicin release and the immunity protein that protects the host from its own colicins (Riley, 1993, Cascales et al., 2007).

Here I sought to identify the mechanisms in *S. sonnei* responsible for interbacterial competition with a relevant microbiota competitor, *Escherichia coli,* and relate this to established epidemiological understandings. To do so, I leveraged sequencing data and isolates (n=164) from a previous genomic epidemiology study of cross-sectional national surveillance the United Kingdom from 2008 – 2014 (Baker et al., 2018b). As I were interested in comparing the behaviour of sublineages, experimental replicates were conducted at an epidemiologically relevant level (i.e. testing many clinical isolates from each sublineage rather than testing fewer representatives multiple times) in an experimental framework herein coined Bulk Phenotyping of Epidemiological Replicates (BPER). I advocate BPER as a concept for moving on from extrapolating from model organisms in the genomic era while also overcoming some of the complications of using clinical isolates for laboratory studies. By coupling BPER to bacterial Genome Wide Association Studies (GWAS), I demonstrate that *E. coli* killing in *S. sonnei* is mediated by colicins found in epidemiologically successful sublineages.

4.2. Materials and methods

4.2.1. Strains and their whole genome sequences

The main collection of *S. sonnei* strains used in this study originated from the archive of isolates from the national reference laboratory of UK Health Security Agency collected from 2008 – 2014 and has been described before(Baker et al., 2018b). These isolates (n = 164) were whole genome sequenced via Illumina HiSeq with 150bp paired end reads. An additional isolate, CIP106374, which was also previously described was included in the bioinformatic

analyses. The assembled and annotated whole genome sequence of this isolates carried out using Oxford Nanopore sequencing as previously described (Anderson et al., 2017b) Sequencing data for all isolates is deposited in the European Nucleotide Archive and individual accession numbers are provided in Supplementary Table 5.

4.2.2. *Phylogenetic tree construction*

FASTQ sequence files for the 161 isolates that were culturable were mapped against the reference *S. sonnei* 2a strain 53G (HE616528.1) concatenated with its multiple plasmids (HE616529.1, HE616530.1, HE616531.1, HE616532.1) using Burrows-Wheeler Aligner (BWA) v0.5.9-r16 (Li and Durbin, 2009). The mapping files were filtered and sorted using samtools (Li et al., 2009b). Duplicates were marked using Picard tool. Subsequent variant calling was completed through bcftools and a consensus file was generated for each isolate (Danecek and McCarthy, 2017). Each chromosome sequence was extracted, and regions were masked using a mask file containing plasmid sequences, IS elements and repeat regions. Gubbins v2.3.4 removed duplicate and low-quality sequences followed by SNP-sites to obtain the core-genome alignment (Page et al., 2016, Croucher et al., 2014). RAxML-ng was then utilised to infer a phylogenetic tree (Kozlov et al., 2019). Each phylogenetic tree has been midpoint rooted with visualisations completed using interactive Tree of Life (iTOL) v6.1.1 (Letunic and Bork, 2019).

4.2.3. *Genotyping of* S. sonnei

Mykrobe v0.10.0 was utilised on the FASTQ sequences of all isolates (Hunt et al., 2019b). The output from Mykrobe was then parsed using a custom python script (https://github.com/katholt/sonneityping/) based on the genotyping scheme proposed by Hawkey *et al*., (2021) (Hawkey et al., 2021). Upon parsing of the Mykrobe output, a tsv file containing the genotype was generated.

4.2.4. *Initial screen of Subclade representatives for killing*

An initial screen of seven strain representatives was screened for killing (Supplementary Figure 3). Pre-cultures of both *S. sonnei* and GFP-producing *E. coli* were grown in tryptone soy broth (TSB) at 37˚C in a shaking incubator (220rpm) overnight. Pre-cultures were then diluted 1:100 (v/v) in fresh TSB and grown until mid-log phase was reached before mixing 1:10 (*E. coli*: *S. sonnei*) adjusted by the OD600 value for the competition mixture. 10µl of the competition mixture was then spotted onto a sterile nitrocellulose filter paper placed on a tryptone soy agar (TSA) plate and incubated at 37˚C overnight without shaking. Competition mixture was then washed off into 500µl of sterile PBS, serially diluted and plated on selective media for CFU enumeration. *E. coli* was selected on 30µg/ml kanamycin and *S. sonnei* was selected on 30µg/ml azithromycin where applicable. Chromosomally encoded GFP of the *E. coli* strain was used to further differentiate *E. coli* from *S. sonnei*.

4.2.5. BPER using cell sorter

S. sonnei isolates (n=161) were grown overnight in 96 well flat bottom plates (Greiner Bio One, UK) containing 150µl of TSB at 37˚C with shaking at 220rpm and diluted 1:100 (v/v) into fresh TSB and grown for two hours in similar conditions. *E. coli* was grown and diluted into fresh media as was done for initial killing assays and 880µl of mid-log phase culture was diluted in 52ml of TSB and 130µl of that was distributed into each well of the 96-well plate. 20µl of the mid-log phase *S. sonnei* cultures were then added to each well to make up a final volume of 150µl with 1:10 (*E. coli*: *S. sonnei*) of the competition mixture. The 96-well plates with the competition mixtures were then incubated at 37˚C with shaking at 220rpm overnight. The overnight competition mixtures were then diluted and GFP expressing *E. coli* cells which had survived the competition with *S. sonnei* were counted using a Bio-Rad ZE5 Cell Analyzer in a total of 10,000 events per well. The percentage of GFP expressing cells were calculated using FCS Express version 7 (De Novo Software) and plotted onto the phylogenetic tree using iTOL (Letunic and Bork, 2021).

4.2.6. BPER growth assays using plate reader

Competition mixtures from overnight pre-cultures were set up in the same way as was done for the BPER assay using the cell sorter, except black 96-well plates (Greiner Bio One, UK) were used to avoid cross contamination of GFP fluorescence signal and the growth of the competition mixture was monitored by growing them in a Synergy H1 multi-mode plate reader (BioTek Instruments Inc) at 37˚C with shaking. Readings for the GFP signal was recorded every 15 minutes during the overnight growth of the competition mixture and *E. coli* killing was determined based on the GFP signal produced from each well where a clear signal of GFP was considered as the inability of *S. sonnei* to kill *E. coli*. A GFP% cut off of 20% was utilised where >20% indicated a non-killing phenotype.

4.2.7. Screening isolates for the T6SS

The 18 components of the full T6SS were extracted from the annotated Nanopore sequence of *S. sonnei* CIP106347 and compiled into a singular multi-FASTA file. SRST2 v0.2.0 (Inouye et al., 2014) was then utilised for the purpose of gene detection using a custom database. The T6SS components within the multi-FASTA were not pre-clustered and therefore were assigned to gene clusters based on 90% nucleotide similarity via CD-HIT v4.8.1 (Li and Godzik, 2006). Python scripts provided as part of the SRST2 package were utilised to parse the clusters and generate a SRST2 compatible database. SRST2 gene detection was then run utilising the custom T6SS database.

4.2.8. Genome Wide Association Study

Draft assemblies of all isolates were assembled using Unicycler v0.4.8 (Wick et al., 2017b) and annotated using Prokka v1.14.6 (Seemann, 2014) Paired end reads were mapped to the reference genome using Burrows-Wheeler Aligner (BWA-MEM) v07.17 (Li and Durbin, 2009) and Picard v2.23.1 was utilised to mark duplicates. Variant calling and subsequent filtering using Freebayes v1.3.2 (Garrison, 2012) was completed and the VCF files generated were merged and used as an input for the GWAS SNP analysis. To generate the input for the GWAS kmer analysis, kmers were counted from assemblies using fsm-lite v1.0. Lastly, to generate the input file for the GWAS COG analysis, Roary v1.007002 was utilised to generate a gene

presence/absence Rtab file containing the presence or absence of each gene in each isolate (Page et al., 2015a).

After generation of appropriate inputs, GWAS was carried out using Pyseer v1.3.6 (Lees et al., 2018). Pyseer uses linear models with fixed or mixed effects to estimate the effect of genetic variation in a bacterial population on a phenotype of interest, while accounting for potential confounding by population structure. For this investigation, presence or absence of the *E. coli* killing phenotype was utilised as the categorical phenotype (Supplementary Table 2). To account for population structure, all analyses were supplemented with phylogenetic distances from the mid-point rooted core genome phylogeny (above). Pyseer analyses were run using the linear mixed model (LMM). Further GWAS investigations were carried out using Scoary v1.6.16 (Brynildsrud et al., 2016a) utilising the gene presence/absence .csv file generated via the Roary pangenome pipeline as well as a trait file consisting of the *E. coli* killing phenotype for each sample.

4.2.9. Colicin database construction and detection

To investigate the presence or absence of specific colicins within the isolates, a large quantity of colicin sequences were collated. Over 10,000 colicins from over 50 species of bacteria were collated from the European Nucleotide Archive as well as including some isolates from previously published sources (Hahn-Lobmann et al., 2019). A multi-FASTA file containing the collated colicin sequences was utilised to generate a custom database via the prepareref command of ARIBA v2.14.6 where prepareref removes erroneous data and runs cd-hit to cluster the sequences based on a user-defined similarity threshold (90% in our case). ARIBA was then run with the FASTQ files of all isolates and the colicin database to report which sequences were observed in each isolate. The database of colicin sequences are publicly available through figshare (10.6084/m9.figshare.20768260.v1).

4.2.10. Statistical testing

All statistical analyses were performed using R v3.6.1. To determine the significant difference between the T6SS profiles of the killing versus the non-killing isolates an independent t-test was undertaken. Further independent t-test were undertaken to determine the relevant significance of the various colicin clusters and determine their order of importance.

4.2.11. Mass spectrometry analysis of supernatants and data post-processing

TCA precipitated proteins from Shigella culture supernatants were resolubilized in UTCT buffer (urea 2M, thiourea 7M, CHAPS 4%, Tris HCl 20 mM pH 7.8) for 1 hour at 30°C under shaking (800rpm) and clarified by centrifugation at 12 000 g at 4°C for 15 minutes. Samples were quantified with Bradford assay. 10µg of proteins were precipitated and digested with sequencing-grade trypsin (Promega, Fitchburg, MA, USA). Each sample was further analyzed by nanoLC-MS/MS on a TripleTOF-5600 mass spectrometer (Sciex, Canada) coupled to a U3000-RSLC system (Thermo-Fisher Scientific, USA) as described previously (Roche et al., 2021). Data were searched with target-decoy strategy against the UniProtKB database from *S.sonnei* (release 2021_01, 22219 sequences) concatenated to a home-made database (details currently being provided but not available at this point) consisting of all *E.coli* colicins, immunity and lysis proteins. Peptides and proteins were identified with Mascot algorithm (version 2.8.1, Matrix Science, London, UK) and data were further imported into Proline v2.1 software (Bouyssié et al, 2020). Proteins were validated on Mascot pretty rank equal to 1 and 1% FDR at peptide spectrum match (PSM score) level. The total number of MS/MS fragmentation spectra was used to quantify each protein in each biological sample and in two technical MS replicates. The total number of MS/MS spectra was computed while considering shared and unique peptides (BASIC Spectral Count) or only unique peptides (SPECIFIC Spectral Count). The mass spectrometric data were deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXDxxxx (to be confirmed on publication).

4.3. Results

During this study a step-wise approach was used to integrate genomic epidemiology, high throughput phenotypic screening, bacterial GWAS, and laboratory validation (Figure 15).

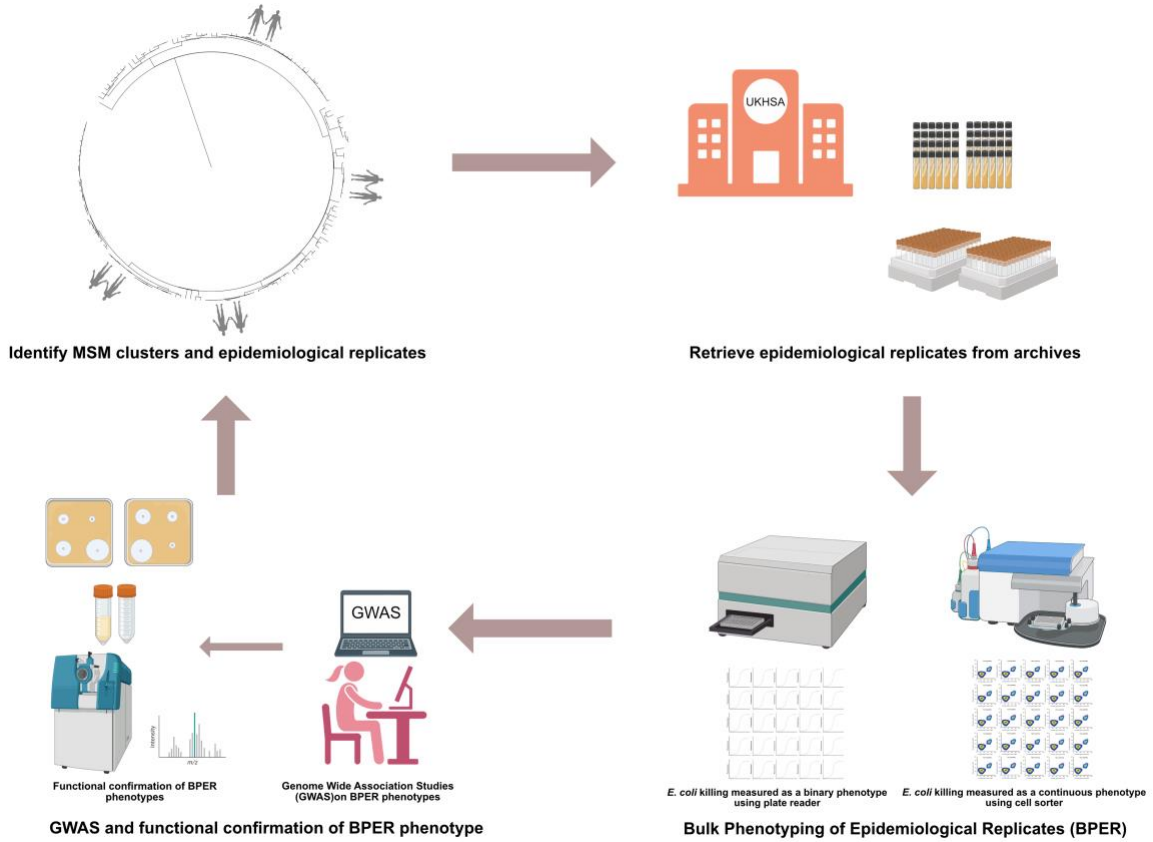


Figure 15: Overview of approach taken within this study.

Schematic overview of the approach used in this study, including Bulk Phenotyping of Epidemiological Replicates (BPER)

4.3.1. The epidemiology and global context of the isolate collection

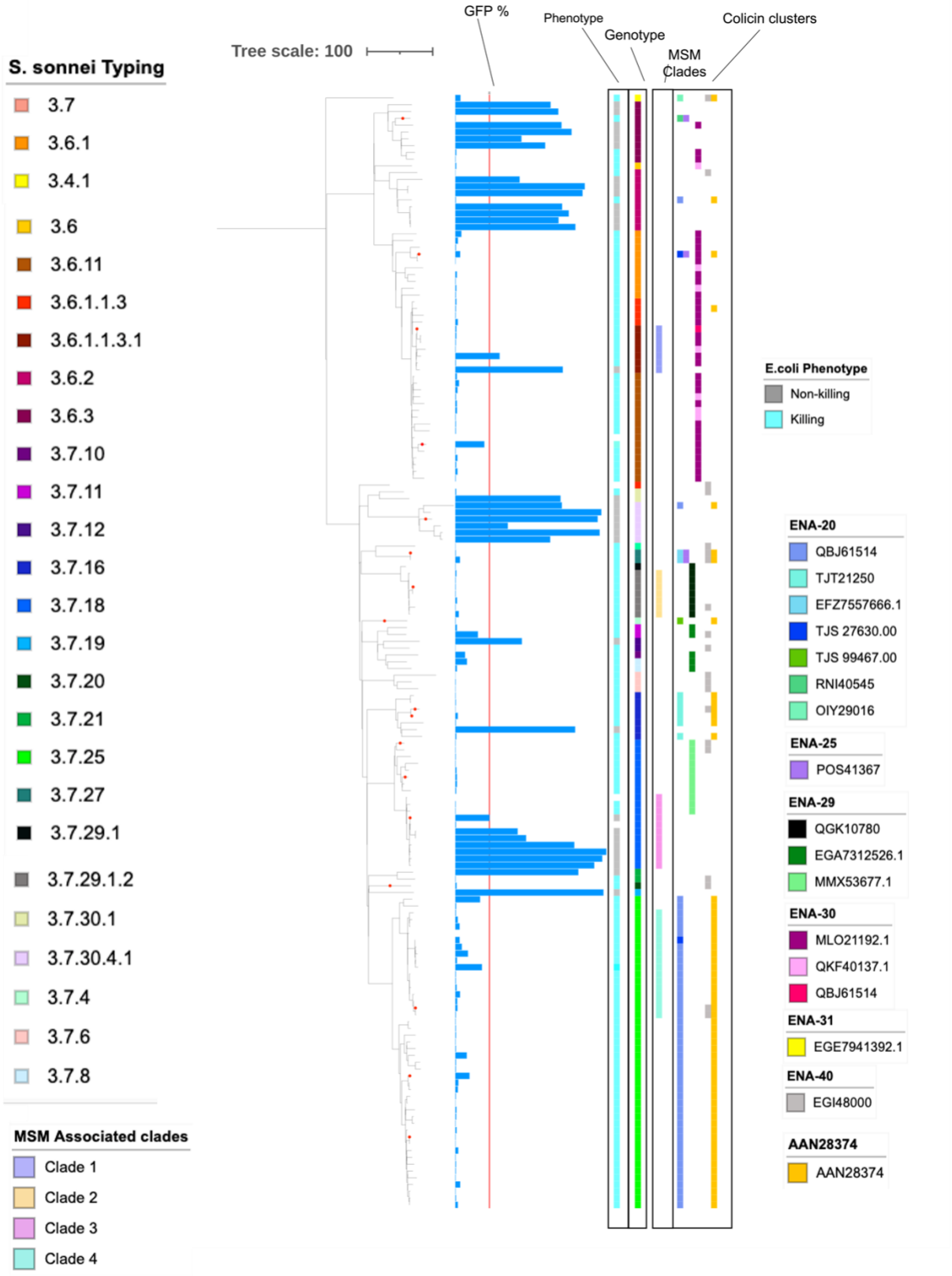


Figure 16: Contextualisation of Lineage 3 *S. sonnei* isolates with respect to phenotype, genotype and key colicin clusters thought to contribute to *E. coli* killing phenotype.

The tree is a midpoint rooted maximum likelihood phylogeny for 165 *S. sonnei* Lineage 3 isolates. Red circles overlaying tree tips indicate isolates which were utilised for mass spectrometry. The horizontal blue bars represent the GFP% with the vertical red line highlighting the 20% cut off where isolates > 20% indicate a non-killing phenotype. The killing/non-killing phenotypes are depicted in the colour strip closest to the tree followed by genotype, isolates within MSM clades and key colicins clusters in the following colour strips. Keys correspond to the relevant colour strips.

A collection of *S. sonnei* isolates used in this study came from a cross sectional subsample of routine microbiological surveillance in the United Kingdom from between 2008 to 2014 (n=164) (Baker et al., 2018b). The original epidemiological study revealed the presence of four distinct clades which were co-circulating among MSM. In this study, Clades 1, 2, and 4 showed markers of population level epidemiological success (specifically international spread and prolonged circulation) relative to Clade 3, which was not successful (Baker et al., 2018b). Notably, a recently described international global genotyping framework has renamed these Clades 1 – 4 (above) as Subclades 3.6.1.1.3.1 (CipR.MSM1), 3.7.29.1.2 (VN2.MSM2), 3.7.18, and 3.7.25 (MSM4) respectively (Hawkey et al., 2021). These isolates were utilised along with CIP106347, which was previously used in studies of a putative *S. sonnei* T6SS (Anderson et al., 2017b) to construct a detailed phylogeny complemented with genotype assignations (Figure 16).

All *S. sonnei* isolates belonged to the globally disseminated multidrug resistant of *S. sonnei*, Lineage 3 (n=165) (Figure 16). Further genotyping showed that most isolates (61%, n=101) belonged to Clade 3.7, particularly Subclade 3.7.25 (MSM4, 28%, n=46) and Subclade 3.7.18 (12%, n=19) (Table 5). The remaining isolates (39%, n=64) belonged to Subclade 3.4.1 (Latin American III) and various Subclades of Clade 3.6 (Central Asia III) including 3.6.1 (CipR_parent), 3.6.1.1 (CipR), 3.6.1.1.3, 3.6.1.1.3.1 (CipR MSM1), 3.6.2, and 3.6.3 (Table 5). Correlating the genotype names with epidemiological history revealed that various globally important subclades of *S. sonnei* were captured in our dataset including internationally

disseminating antimicrobial resistant genotypes (Table 5 and Supplementary Table 5). Thus, our collection contained a breadth of the diversity of the globally disseminated Lineage 3. All metadata and sequence data accession numbers are provided in Supplementary Table 5.

Table 5: Population structure summary of *S. sonnei* genotypes in this study^

Genotype	Isolates (N)	Name	Original Name	Epidemiological summary	Reference
3.6.1	10	CipR_parent	-	Subclade from which ciprofloxacin-resistant sublineage emerged	(Hawkey et al., 2021)
3.6.1.1	16	CipR	Ciprofloxacin-resistant Pop2	Triple QRDR* mutation ciprofloxacin-resistant sublineage	(Chung The et al., 2019, Chung The et al., 2015)
3.6.1.1.3	4	-	-	Ciprofloxacin-resistant	(Baker et al., 2018a, Chung The et al., 2019)
3.6.1.1.3.1	7	CipR MSM1	MSM Clade 1	MSM-linked ciprofloxacin resistant isolates	(Baker et al., 2018a)
3.6.2	9	Central Asia III Subclade	-	Associated with areas in Central Asia	(Baker et al., 2018a, Baker et al., 2016)
3.6.3	9	Central Asia III Subclade	-	Associated with areas in Central Asia	(Baker et al., 2018a, Chung The et al., 2019)
3.7.16	7	Global III Subclade	-	-	(Holt et al., 2012b)
3.7.18	19	Global III Subclade	MSM Clade 3	MSM-associated	(Baker et al., 2017)
3.7.25	46	MSM4	MSM Clade 4	MSM-associated	(Baker et al., 2018a)
3.7.29.1.2	7	VN2 MSM2	MSM Clade 2	MSM-associated. Emerging from sweep 2 of Vietnam clone	(Bardsley et al., 2020, Holt et al., 2013)
3.7.30.4.1					

	6	OJC	OJC-associated	Associated with the Orthodox Jewish communities in Israel, UK, USA and Europe	(Baker et al., 2016)

* QRDR = quinolone resistance determining region

^ for brevity, only genotypes that included >4% of isolates are included in the table. Full genotyping data is available in Supplementary table 5.

Table 6: Summary of bacterial strains used in this study and their origins.

Strains	Isolates (N)	Description	Reference
***S. sonnei*^**	164	Cross-sectional isolates from routine surveillance during 2008-2014 collected by Public Health England (now UKHSA).	(Baker et al., 2018c)
***S. sonnei* CIP106347**	1	*S. sonnei* clinical isolate from the collection at Institut Pasteur, previously described as having a functional T6SS	(Anderson et al., 2017a)
***E. coli* MG1655**	1	*E. coli* MG1655 strain with a chromosomally encoded, constitutively expressed GFP marker and a kanamycin resistance marker	Kind gift from Bottery Lab (University of Manchester), used in (Malaka De Silva et al., 2022)

^ Various – see supplementary Table 5

4.3.2. E. coli *killing is common and associated with genotype in* S. sonnei

After an initial low throughput screen revealed differences in the *E. coli* killing phenotype among Subclade representatives (Supplementary Figure 3), it was decided to scale our approach to include the entire collection of *S. sonnei*. This was done so our laboratory assays

were being replicated out at our desired level of inference i.e. I wanted to relate the phenotype to the behaviour of different genomic subtypes, so I needed multiple clinical isolates that belonged to those subtypes, rather than biological or technical replicates (though for robustness it was chosen to also included duplicate technical replicates), which is coined BPER. Three isolates from the collection failed to grow in the laboratory so BPER assays were conducted with 161 isolates.

Owing to the number of isolates involved and our ambition to scale the BPER approach further in future studies, assays of the killing phenotype were carried out in a high-throughput manner. Specifically, cell sorting was used to measure the proportion (%) of green-fluorescent *E. coli* remaining after overnight competition with *S. sonnei* in a 96-well format. As a cell sorter is specialist equipment, the results were correlated with a binary read out of green fluorescence following a simple plate-reader based growth assay with greater availability across laboratories. Introducing a threshold of positivity among the cell sorting results of these assays led to 99% concordance between the two approaches (Figure 16, Supplementary Table 1, see methods). It was found that most 81.36% (131/161) of our Lineage 3 *S. sonnei* displayed *E. coli* killing with phenotype clustering by genotype (Figure 16). The continuous data (from cell sorting) revealed a spectrum of difference in the phenotype with the proportion of *E. coli* following competition being between 0% – 88.54% (mean = 13.85%). Associating the phenotype against the population structure of *S. sonnei* revealed that Subclades of the Central Asia III Clade (3.6.3 and 3.6.2), 3.7.18 (MSM Clade 3), and 3.7.30.4.1 (a Subclade associated with transmission among Orthodox Jewish Communities) were predominately non-killing with the remaining Subclades being predominately killing (Figure 16). However, differences among isolates belonging to the same Subclade (i.e. epidemiological replicates) demonstrated phenotypic variation (e.g. genotype 3.7.18 where

isolates ranged from 0.1% – 88.44% with a mean of 25.15%), highlighting the value of the BPER approach when assaying clinical isolates for inference at an epidemiological level.

4.3.3. *GWAS indicates that colicins are responsible for* E. coli *killing in* S. sonnei

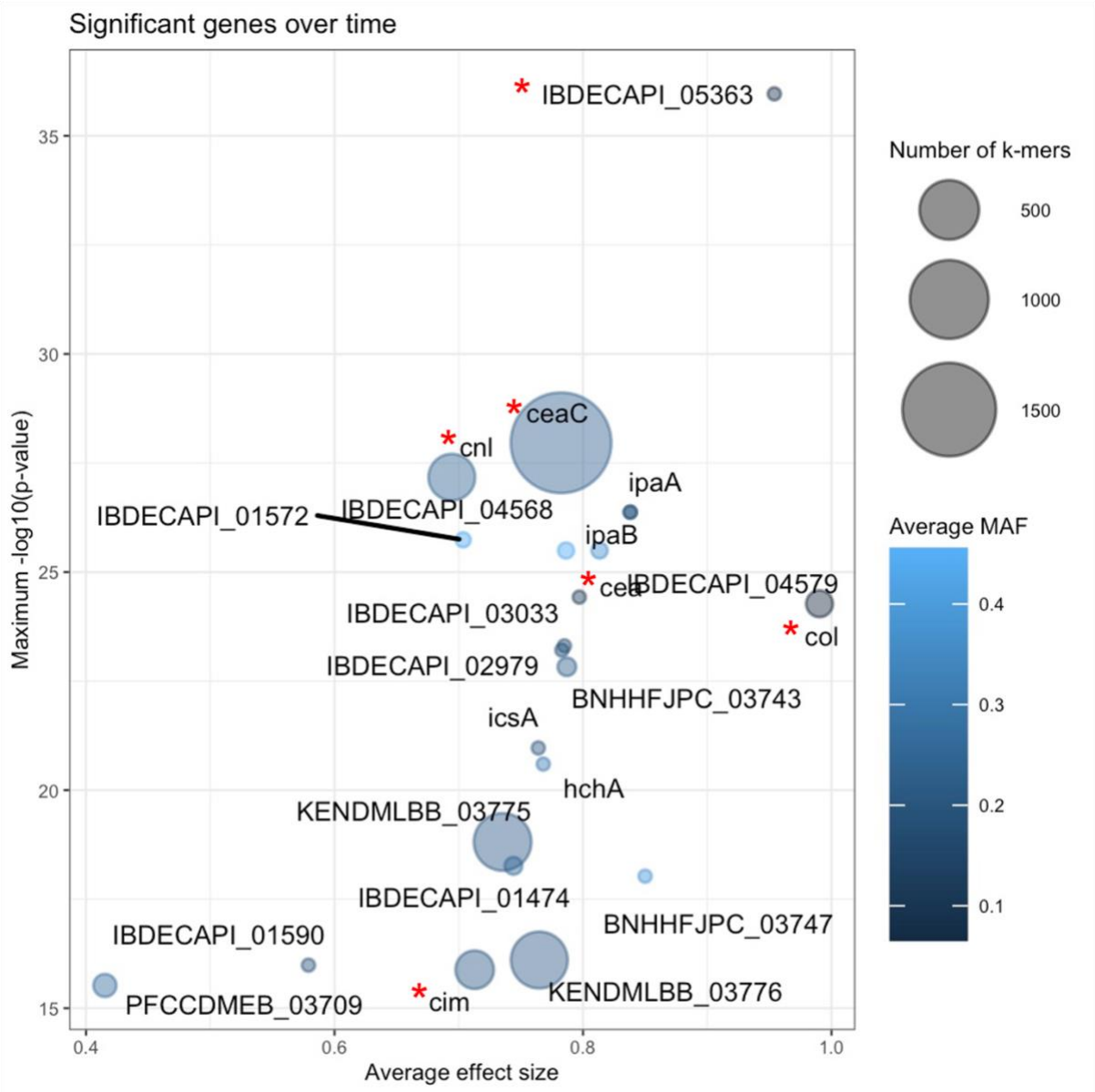


Figure 17: Genetic feature association in GWAS by kmers.

Bubble plot showing the number of kmers (bubble size) by gene (text labels in graph field), effect size (x-axis), and statistical support (y-axis as negative logarithm of the p-value). Red stars indicate those with colicin related functions.

To identify genetic factors responsible for *E. coli* killing, I conducted a bacterial GWAS for association with the *E. coli* killing phenotype (using pyseer, see methods) and focused on

those results that related to predicted/annotated genes. This revealed a variety of genetic factors that were significantly positively associated with the *E. coli* killing phenotype including 3187 kmers which were within 22 genes (short sequence fragments of length k=10-100, maxp>15) and 64 clustered orthologous groups (COGs) (lrt p-value<0.05) (Supplementary Table 6).

Investigating the 22 genes which contained significantly associated kmers revealed 1 hypothetical gene and 21 of known function; 6 of which were colicin related. These were IBDECAPI_05363, *cim*, *cnl*, *cea*, *ceaC* and *col* (Figure 17, Supplementary Table 6). IBDECAPI_05363, which was the most significant result with the largest effect size (Figure 17), displayed 99% similarity to the *E. coli* E1 colicin immunity protein despite not being annotated in *S. sonnei*. The former, *cim*, originally described on the *E. coli* plasmid CloDF13 is known to be bacteriocinogenic and *cnl* encodes a lysis protein for colicin N originally described as encoded on the *E. coli* plasmid pCHAP4 (Pugsley, 1988, Nijkamp et al., 1986). The genes *cea*, *col* and *ceaC* encode for the colicins E1, E2 and E3 respectively and were originally identified on small plasmids ColE1, ColE2 and ColE3 (Bishop and Hunt, 1988, Watson et al., 1981). In further support of the contribution of colicins to the killing phenotype, two of these genes (*cnl* and *ceaC*) were the first and ninth best supported genes in the COG analysis, and a COG analysis using an alternative GWAS approach (scoary, see methods) also identified the colicin related genes *imm* and *cnl* as strongly associated with the *E. coli* killing phenotype (Supplementary Table 6). Owing to the significant burden of evidence that colicins were responsible for the *E. coli* killing phenotype, I explored the distribution of these proteins further (below).

Genes containing kmers positively associated with the killing phenotype that were not obviously related to colicins included phage related genes (n=4), T3SS related genes (n=3),

DNA modification/binding genes (n=2), protein deglycase gene (n=1) and also plasmid replication/transfer genes (n=5) (Supplementary Table 6). Notably, the plasmid replication/transfer genes from the kmer analysis (KENDMLBB_03776, BNHHFJPC_03747, PFCCDMEB_03709, BNHHFJPC_03743 and IBDECAPI_04568) encoded for replication protein B (IBDECAPI_04568), repE (BNHHFJPC_03743) with the latter (RepE protein) appeared as an intersecting result with the COG analysis results. Replication protein B and RepE have been shown to initiate replication of both chromosomal segments and plasmids within *E. coli* highlighting its potential for initiation of small plasmids such as those encoding colicins (Miller and Cohen, 1999). It is possible that these plasmid-related genes were associated due to the importance of the colicin-related genes and so are within close proximity leading to their co-occurrence in GWAS.

4.3.4. *Various colicins are widely distributed in* S. sonnei

To determine the distribution of colicins in *S. sonnei* I screened the genomes against a custom database of >10,000 colicin sequences that grouped into 145 clusters (methods). Although this revealed that widespread distribution of colicins (all isolates contained ≥1 colicin), no single colicin or cluster showed perfect concordance with the presence of the killing phenotype across *S. sonnei* (Supplementary Table 5), suggesting no single protein was responsible for the phenotype. To determine which colicin clusters were most likely responsible for the killing phenotype, I determined the association of colicin clusters with the binary *E. coli* killing phenotype (methods). The clusters were ordered based on significance (most significant first) and each cluster was added in this order until all *E. coli* killing could be explained without clusters overlapping significantly with the non-killing phenotype (Figure 16). until all killing phenotypes were explained. This process resulted in seven key clusters being identified that explain the distribution of the *E. coli* killing in *S. sonnei*.

To further validate the association of colicins with the *E. coli* killing phenotype, 15 isolates representative of the breadth of the key colicin clusters distributed throughout the phylogeny and 2 isolates that did not display *E. coli* killing as controls were selected (Figure 16). Extracted culture supernatants of these isolates (n = 17) were used to confirm the presence of the colicins in the supernatant. Laboratory assay for contact-independent killing confirmed that filtered supernatants from the 15 colicin cluster-containing *S. sonnei* killed *E. coli* while those from the two non-killing isolates did not (Supplementary Figure 4). Mass spectrometry on filtered supernatants from all isolates (100%, 15 of 15) that displayed *E. coli* killing contained at least one contributing colicin in the total peptides identified in the samples and, for 87% (13 of 15) of isolates, the proteins matched to a unique peptide sequence of an individual colicin (Supplementary Table 7). Mass spectrometry also confirmed that, the supernatants from the two isolates that did not display *E. coli* killing did not yield any matches to any of the colicin sequences further confirming the role of colicins in *E. coli* killing (Supplementary Table 7).

4.3.5. E. coli *killing in vitro in* S. sonnei *is not mediated by T6SS*

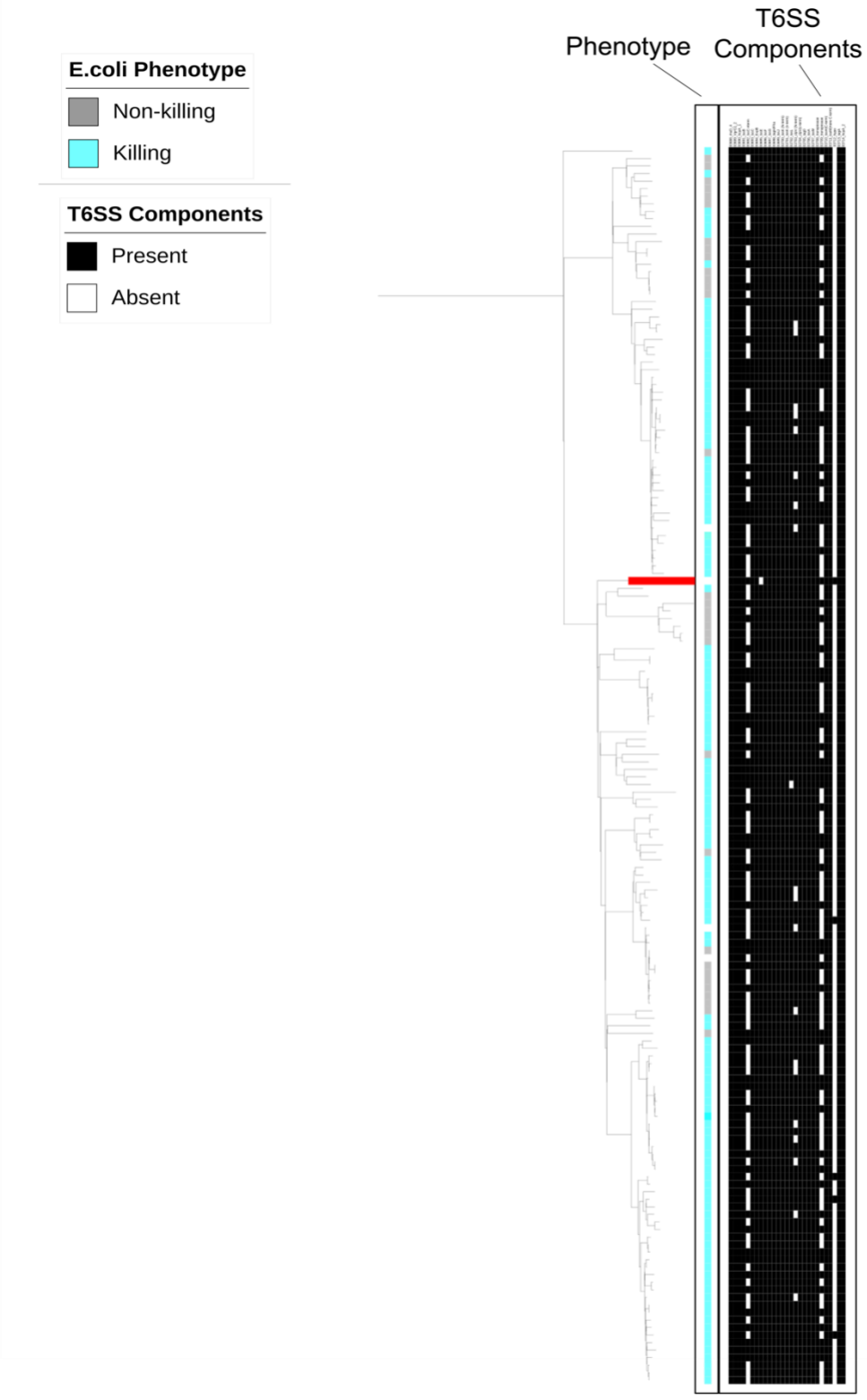


Figure 18: T6SS profiles for Lineage 3 *S. sonnei* isolates.

The tree is a midpoint rooted maximum likelihood phylogeny for 165 *S. sonnei* Lineage 3 isolates. Red colour block highlights the CIP106347 isolate. The killing/non-killing phenotype is depicted in the colour strip closest to the tree followed by colour strips of the presence/absence of the various T6SS components to display the T6SS profiles of each isolate.

Our results highlight the critical role of colicins in mediating *in vitro E. coli* killing in *S. sonnei*. As a previous study described a functional T6SS in *S. sonnei* clinical isolate CIP106347 that was hypothesised to contribute to competition with *E. coli* and *S. flexneri* I also explored the possible role of the putative T6SS (despite none of the components being identified by GWAS) (Anderson et al., 2017b). To do this CIP106347 was re-sequenced and I extracted the gene sequences for predicted proteins in the region of the T6SS system (predicted proteins were used as the reconstituted locus was inconsistent with the originally described schematic) (Anderson et al., 2017b). In fact, one or more of the key components of the T6SS (i.e., transposase or N terminus of TssC) were absent in 100% of our isolates and in CIP106347 suggesting that the T6SS in *S. sonnei* is likely non-functional (Figure 18). Consistent with this suggestion is that lack of a correlation between the presence and absence of components of the CIP106347 T6SS locus and the killing phenotype in our studies (Figure 18). To further support this, the difference between the number of T6SS elements per isolate for the killing phenotype ($\bar{x}$=16.27, n=133) and the non-killing isolates ($\bar{x}$=16.28, n=36) was not statistically supported to differ (t_{169}, = -0.059, P = 0.9527). In conclusion, regardless of the functionality of the T6SS, the profiles did not show any significant difference between the *S. sonnei* isolates in our collection that did and did not exhibit *E. coli* killing, so the T6SS is not responsible for our phenotype.

4.4. Discussion

Outcompeting commensals like *E. coli* in the gut is an important aspect for *Shigella* for surviving and establishing in their niche. Since *Shigella* is known to have a relatively lower infectious dose, I can expect that the mechanisms by which *Shigella* establishes itself are relatively efficient (Porter et al., 2013). Therefore, it is important to investigate these mechanisms to better understand the dynamics of the complex gut microbial communities that lead to infection, and correlate this with established epidemiological understandings. While there will be several mechanisms at play in the complex environments and microbial communities that the pathogens live in, I demonstrated here the crucial role colicins play in interbacterial competition with *E. coli* and the consequences this contributes to at an epidemiological level (Lindsay et al., 2015, Pop et al., 2014). The important role of colicins in *Shigella* as a genus could be further evidenced by recent reports of colicins being discovered in species such as *S. flexneri* 2a (Torrez Lamberti et al., 2022) and a previous description of colicin plasmid acquisition in endemic *S. sonnei* in Vietnam (Holt et al., 2013) Our findings strengthen this evidence base by our use of diverse collection of real-world isolates and assay of the phenotype with replication at an epidemiologically relevant level through BPER.

The high prevalence of *E. coli* killing throughout the *S. sonnei* phylogeny suggests that this is an epidemiologically important phenotype. This is further supported by the only non-killing Subclade (3.7.18, MSM-associated Clade 3) being outcompeted by *E. coli* killing Subclades (3.6.1.1.3.1, 3.7.25, and 3.7.29.1.2) in a scenario where the subclades were known to be co-circulating in a single patient population (Baker et al., 2018c). Of course, phenotypes other than *E. coli* killing will also contribute to the competition dynamics among different genotypes. Indeed, our previous work identified that 3.7.18 was also comparatively distinct in lacking a low fitness azithromycin resistance plasmid, pKSR100 (Baker et al., 2018c, Malaka De Silva et al., 2022). Non-killing phenotypes were also conserved in other Subclades of

Central Asia III (3.6.2 and 3.6.3) and 3.7.40.1 (Global III OJC), highlighting the non-necessity of *E. coli* killing. For the majority of Lineage 3 clades the *E. coli* killing ability dominated and was conserved through clade expansion indicating its potential to contribute to the global success of Lineage 3 *S. sonnei*.

The BPER approach is advocated as a way of assessing the phenotypes in an epidemiologically relevant way. Since most studies carry out laboratory experiments of phenotypes using either type strains or a limited number of clinical isolates, the level of applicability of what could occur in a real-world setting might not be fully captured. Therefore, it is plausible propose that using a collection of real-world isolates in laboratory phenotype experiments and having epidemiological replicates where possible rather than biological replicates offer greater insight into relevant pathogen biology. It was found that, in the case of this phenotype, BPER could be implemented in a simple, cost-effective experimental setup with no/minimal compromise on accuracy (only two mismatches between the experiments using flow cytometry and plate reader, Supplementary Table 2). Here, by working with a real-world isolate collection with known epidemiological outcomes I could work backwards to identify the factors that helped shape the observed epidemiology and deepen our understanding of the biology in a targeted, comparatively rapid manner.

I have shown that *in vitro E. coli* killing by *S. sonnei* is common and mediated by colicins, not a T6SS as previously suggested (Anderson et al., 2017a). This was supported by the lack of a fully intact T6SS in any of the isolates under study and no correlation of putative T6SS components and the killing phenotype. Furthermore, well supported associations between colicin clusters and the *E. coli* killing phenotype *in silico* and *in vitro* confirmation of the secretion and functionality of colicins demonstrates that colicins are responsible for *E. coli* killing by *S. sonnei*. Excitingly, our work also identified several novel genes that were

associated with the *E. coli* killing phenotype not known to be part of colicin synthesis or activity (i.e. four phage related genes and one hypothetical gene). These may have some currently unknown relationship with the colicins (e.g. phage proteins may be involved in the mobility of the colicin plasmids) or have as yet unknown functions. In any case, these genes represent ideal candidates for future studies investigating the mechanisms of interbacterial competition in *Shigella*. The demonstrated importance of interbacterial competition in our epidemiological setting raises the question of the importance of this phenotype in shaping other bacterial population dynamics where this should be further investigated.

Chapter 5

5. Discussion

The purpose of this thesis was to identify key factors which are contributing to the success of *Shigella* species as a global public health pathogen. Through novel and traditional GWAS strategies key factors for success over time for both *S. sonnei* and *S. flexneri* were identified as well as potentially identifying the cause of a key *E. coli* killing ability within modern *S. sonnei* Lineage 3. Below I will briefly consolidate and summarise key findings based upon key areas of this work such as the utility of historical isolates and the validity of tGWAS as a methodology. Notably, suggestions for improvement and caveats for each investigation are discussed.

I then finalise this thesis by contextualising these studies in the broader field of success in *Shigella* species as well as other bacterial taxa whilst commenting on future directions and implications of this work.

5.1. The utility of historical isolates

5.1.1. The advantages

Historical isolates have been an underused resource in previous years. I think that the underuse of historical isolates first stemmed from the difficulties in sampling, contamination, and low concentrations of DNA. The first bacterial genome sequence, a *Haemophilus influenzae* isolate, was only published in 1955 which is much later than many historical isolates originate from (Fleischmann et al., 1995). Obviously, sequencing and sampling strategies have improved substantially in recent years making sequencing cost-effective, accurate and routine in some cases. Now that difficulties such as contamination and low quantities of DNA can be overcome, historical isolates are more accessible. However, I do

believe that in general there has been an increasing interest in being able to fully understand the evolutionary arc of pathogens to better facilitate understanding of virulence, resistance, and project future trends.

Here, I have utilised the historical Murray Collection to elucidate on the successful evolution of *Shigella* as a public health pathogen. Contextualisation of both *S. sonnei* and *S. flexneri* historical isolates within modern population structures confirmed previous biological understanding with respect to how each species population is affected by emerging PGs or Lineages. Furthermore, the historical AMR profiles of the *Shigella* isolates demonstrated low quantities (n=1/2) or a complete lack of AMR determinants. Small quantities of AMR determinants again provides evidentiary support for previous biological understanding for the ancient origins of certain AMR determinants (e.g *blaEC-8*) to confer resistance to antibiotics present in the environment (Barlow and Hall, 2002). It truly is a testament to the vital nature of selective pressure, in this case discovery and use of clinical antibiotics, to drive acquisition and stabilisation of AMR determinants. Here the utilisation of historical isolates has confirmed biological understanding of population structure and AMR evolution, highlighting their impressive and rich nature as an underused resource.

The primary advantage of incorporating historical isolates with the plethora of readily available modern isolates in this study was to increase the time span to be inclusive of key time periods e.g the pre-antibiotic era. For the purposes of tGWAS where its primary function is to identify factors positively associated with time, it would clearly be beneficial to have as long a time span as conceivable to characterise as many positive associations as possible over time. I believe it is plausible that early, and potentially key, drivers of initial success would be missed without the inclusion of historical isolates. Understanding fully the path to success for public health pathogens can aid future management and treatment of many enteric bacteria.

5.1.2. *The potential challenges*

The overarching advantage of historical isolates is to facilitate investigations of unique time periods and events such as the pre-antibiotic era and world wars, not covered by modern WGS surveillance. However, even with improvements in sampling strategies and sequencing technologies there are still some disadvantages to consider when utilising historical isolates. Historical isolates are not as commonplace as modern isolates and so there are a limited number of historical isolates available as public resources. Historical isolates are not systematically sampled and so are not equally representative for each year, with some years having none, leading to them being potentially unrepresentative of the entire time-period, as demonstrated by >80% of the *S. sonnei* Murray collection isolates being from 1937. This allows bias to be introduced into the data. Many historical isolates also lack much of the modern metadata typically taken for granted, including geographical location. The lack of metadata means that any epidemiological inferences cannot be made even when used in tandem with modern isolate collections where geography is included.

Overall, I believe this study has showed that even with the possible challenges posed by historical isolates they can be a valuable addition to studies investigating bacterial evolution over time. They offer a unique and rich insight into time periods previously under-characterised and allow a full path to success for public health pathogens to be mapped.

5.2. The rabbit hole of GWAS approaches and the validity of tGWAS

5.2.1. *GWAS: The future of genomics?*

The remarkable range of discoveries that GWAS have facilitated in genomics has been astounding. In a review released in 2021 it was noted that more than 5700 GWAS have been conducted for over 3300 traits yielding understanding into a variety of complex diseases and

phenotypes (Uffelmann et al., 2021). Since the first bacterial GWAS was published in 2013, the field has matured substantially with an ever-growing number of tools, traits and isolate collections to fully decipher the complex underlying biology of complex traits related to health and disease (Loos, 2020). The field shows no signs of slowing down and neither should it. There have been ground-breaking insights gained via bacterial GWAS and with the more sophisticated technologies and analytical tools being developed it can only go from strength to strength.

5.2.2. The enigma of bacterial GWAS workflows and analytical pitfalls

One of the points of conjecture for bacterial GWAS methodologies in general is the lack of universal workflows and statistical cut-offs for associated hits. The growing numbers of tools, methods and workflows presents a major challenge to researchers as there is a lack of comparative studies to determine the overlap or discrepancies between tools. There are multiple tools for completing bacterial GWAS including Pyseer, Scoary, treeWAS and PLINK, to name a few (Lees et al., 2018, Brynildsrud et al., 2016b, Purcell et al., 2007, Collins and Didelot, 2018). Each of these bioinformatic tools utilise slightly different GWAS approaches to reveal associations to phenotypes of interest. PLINK represents the traditional SNP-based GWAS methodologies whereas Pyseer, Scoary and treeWAS approaches vary from mixed linear models, pangenome-based or phylogenetic tree-based approaches respectively (Purcell et al., 2007, Lees et al., 2018, Brynildsrud et al., 2016b, Collins and Didelot, 2018). Each tool addresses certain bottlenecks in bacterial GWAS but not others including epistasis, recombination rate adjustments and polygenicity as well as some GWAS tools being restricted to binary phenotypes (e.g Scoary) (Brynildsrud et al., 2016b). All of these distinctive traits make deciding upon a GWAS tool confusing and sometimes leads you down a deep rabbit

hole of complex theoretical GWAS algorithms and adjustments. With the rapid development of new bacterial GWAS tools in today's scientific field it is important to compare these tools and identify overlap and discrepancies. Today, there are now some review articles addressing these issues between different platforms and analytical problems aiding researchers to choose the right platform for their dataset and investigation (Saber and Shapiro, 2020, San et al., 2019). This is a great next step for bacterial GWAS to become a more defined and universal protocol for bacterial genomics.

For analyses completed during my thesis I chose to utilise Pyseer for the majority of GWAS analyses and Scoary as a comparator for associations contributing to advantageous *E. coli* killing ability within *S. sonnei* Lineage 3. For the tGWAS approach, time was a continuous phenotype and so this eliminated GWAS tools like scoary. Pyseer was ultimately chosen due its capacity to identify short variation (SNPs) as well as longer variation (kmers and COGs) whilst also accounting recombination rates (important for *Shigella*) and population structure. Time as a variable is not as straightforward as a binary trait and so I believed that accounting for multiple types of variation would be the best approach to fully cover positive associations over time. For the *E. Coli* killing GWAS analyses Pyseer was once again utilised, but Scoary was also used to act as a cross comparison to truly highlight what factors were responsible for this phenotype, in this case colicins.

The differences between bacterial GWAS approaches obviously have a big impact on investigation conclusions. Due to these differences in approach, each GWAS methodology will conclude different associations to a phenotype of interest, although there may be some overlap. This was demonstrated in Chapter 4 when I completed both Pyseer and Scoary GWAS approaches to identify associations to the *S. sonnei* Lineage 3 *E. coli* killing ability. GWAS methodologies could only be compared during the COG analyses due to scoary only looking

at gene presence/absence as an input, however, both GWAS approaches resulted in different associations. There was some overlap, especially during the top 10 hits for each analysis, but there were definitive differences noted. This works for and against me for this type of investigation. The associations that were present in both GWAS approaches indicate very strong evidentiary support for their importance to the phenotype and so helps me to decide what could be contributing. However, the question then arises about which other hits to pursue and in what order. There is no correct answer to that question as both are valid GWAS approaches, however, I believe that Pyseer has the advantage of making recombination adjustments and supplies more information on potential contributing factors such as by shorter variation (SNPs). Hence why Pyseer is used as the predominant GWAS approach for my thesis.

Another point of conjecture is the lack of universal cut-offs for statistical significance for any GWAS approach. It is up to each individual researcher to decide reasonable cut-offs for significance which means that there could discrepancies between what was considered significant in different studies so care would need to be taken to truly understand the term significant in each paper. Typically, a p-value of <0.05 would be satisfactory for an association to be considered statistically significant. However, as depicted in the *S. sonnei* tGWAS SNP analysis in Chapter 2 a p-value of <0.05 still qualified >2500 SNPs to be significant, far too many to be feasibly investigated at one time. Therefore, it seemed prudent to raise the cut-off and narrow down the significant associations.

For the tool Pyseer, the documentation stipulates that significance cut-offs for the SNPs and COGs are to be decided by the researcher, whilst they do provide guidance for a significance threshold for the kmer GWAS feature type (Lees et al., 2018). This threshold calculates a p-value threshold based on the Bonferroni calculation and I found this extremely useful to home

in on those kmers within genes associated with the phenotype of interest. Other GWAS tools do not provide scripts to calculate thresholds, but it could be something that develops in the future to help aid a universal understanding of significance in GWAS. For any researcher it is a balance between excluding potentially important associations whilst also managing a feasible quantity of results to investigate. I took the decision to raise certain threshold values such as the significant SNP threshold in both Chapter 2 and Chapter 3 in order keep the results to a manageable level that could feasibly be investigated in the scope of this thesis whilst encompassing enough results to truly answer the scientific questions being asked. This is an ambiguous area for all researchers utilising GWAS tools and future development or guidance in this area would be immensely beneficial, however, I do note the extraordinary task that would be to implement whilst encompassing each studies individual differences.

5.2.3. The validity of tGWAS

In *Shigella*, and in other bacterial taxa, there have been multiple studies highlighting the increasing trend of AMR and, less studied, virulence (Baker et al., 2015c, Holt et al., 2012b, Connor et al., 2015). These studies represented the evidentiary support required to assign these determinants as positive controls to validate tGWAS as an approach. Within both the *S. sonnei* tGWAS analyses in Chapter 2 and the *S. flexneri* tGWAS analyses in Chapter 3, AMR and virulence genes were observed in different GWAS feature types providing substantial support for the validity of tGWAS. Overall, the variety and quantity of AMR and virulence determinants provided more than enough supporting evidence for validity. However, during the *S. sonnei* tGWAS analyses in Chapter 2, I noted that there were smaller quantities of AMR determinants being observed. The only AMR determinant observed was a SNP in *eptB*. The importance of AMR acquisition for success in *S. sonnei* has been well documented especially for lineage 3, the globally dominating lineage (Holt et al., 2012b). The success of this lineage

is largely thought to be due to the acquisition of key AMR determinants such as the Tn7/Int2 cassette (Holt et al., 2012b). Due to the vital nature of these determinants and determinants like these, it may have been surprising to not see more AMR related genes positively associated with time for *S. sonnei*. However, this actually showed that tGWAS was working perfectly. As part of GWAS I adjusted for a set of covariates including lineage to increase precision and remove confounding thereby reducing residual variation and increasing the power to detect an association only with time. These AMR determinants may be strongly associated with a specific lineage and thereby reduced when the covariate file was introduced. tGWAS adjustments successfully accounted for population structure and so therefore it was hard to observe AMR determinants partially due to clonal replacement demonstrated by *S. sonnei*.

Overall, I believe that with the observation of both AMR and virulence determinants throughout both species provides ample evidence as positive controls and also intuitively beneficial genes observed throughout tGWAS to confirm its validity as an approach to investigate bacterial evolution over time. Care must be taken, however, to ensure adjustments are made to account for population structure and covariates to increase the power to detect an association solely with time.

5.3. Insights gained into *S. sonnei*

tGWAS within *S. sonnei* in Chapter 2 revealed a variety positively associated functional genes over time. Apart from AMR and virulence determinants, two main areas of function positively associated over time occurred, namely iron uptake and metabolic mechanisms. Both of these have intuitively beneficial biological roles within *S. sonnei* as well as for other bacterial taxa highlighting the validity and strength of tGWAS.

5.3.1. The importance of iron uptake within human host adapted pathogens

Iron uptake is well documented as important for bacteria, especially intracellular pathogens in human hosts (Caza and Kronstad, 2013). The exceedingly low levels of iron within the human host is actually part of the innate immune system to inhibit invading pathogens (Wei and Murphy, 2016). tGWAS in *S. sonnei* revealed the importance of multiple iron uptake genes (e.g *fec/fes* genes) positively associated over time. This suggests that over time *S. sonnei* has evolved to enhance iron acquisition and hence survival and replication within the human host. The importance of iron acquisition systems is actually two fold – firstly they contribute to their ability to grow and replicate in diverse environments but also they play an important role in the regulation of virulence and metabolic processes (Payne et al., 2006). The iron acquisition systems are intuitively beneficial for the success of *S. sonnei* as a pathogen and have been well documented as essential for any intracellular pathogen within the human host. I believe that as *Shigella* has adapted to the host environment, iron acquisition has gained importance for the bacteria leading to the acquisition, stabilisation, and advantageous mutation of more related genes.

Iron acquisition is essential for a plethora of enteric bacteria. The hypothesis that iron metabolism has gained importance and is central to bacterial success has also been noted and is further supported in other enteric bacteria such as *K. pneumoniae.* Pangenome GWAS (PGWAS) for association with infection in humans revealed siderophore and iron-metabolism genes (Holt et al., 2015). Association of these iron-related genes were highly prevalent within community acquired invasive infections (75% carried one or more) reveals their potential importance to invasive disease as well as their advantageous ability to aid bacterial growth and replication in competitive environments (Holt et al., 2015). Iron metabolism clearly is a

critical part of bacterial pathogenicity and success. A hypothesis supported and validated by tGWAS.

5.3.2. Are bacteria just hungry?

There were a wide variety of catabolic related genes aiding catabolism of varying substrates from citrate to fatty acids to anthranilate. The abundance of catabolism related genes is interesting. It has been previously noted that bacterial metabolism shapes host pathogen interactions responsible for the survival, replication and colonisation of the host via exploitation of the host rich source of nutrients (Passalacqua et al., 2016). It truly is an intrinsic part of bacterial pathogenicity, but one which I believe is often overlooked in favour of more obvious trends such as AMR and virulence. However, there is a growing set of research dedicated to highlighting the overarching importance of catabolism and metabolic changes during infection to aid successful infection supporting my belief that exploitation of the host via catabolic mechanisms may be the main aim of developing bacterial virulence in the first place (Passalacqua et al., 2016, Nogales and Garmendia, 2022).

There is research to suggest the comparable or equivalent importance of AMR and metabolism. In bacterial persisters (bacteria which exhibit extraordinary tolerances to antibiotics), metabolism plays a vital role in the persister phenotype due to its participation in entry, maintenance and exit (Amato et al., 2014). Changes in metabolism to establish a dormant, tolerant state during antibiotic stress and then subsequent metabolic changes on reawakening upon an antibiotic free state make up the backbone of the medically important persister phenotype (Amato et al., 2014). Similarly metabolic states have been shown to be vital for virulence state within other enteric bacteria such as *E. coli* in urinary tract infections (UTIs) (Conover et al., 2016, Nogales and Garmendia, 2022). Studies highlighted the upregulation of non-traditional energy metabolites (e.g non-glucose carbon metabolites) to

aid survival, replication and colonisation under biofilm-like intracellular bacterial communities (IBCs) under oxidative stress (Conover et al., 2016). Without the metabolic changes, virulence determinants alone would not be sufficient to result in successful infection.

There are clearly bodies of research to suggest catabolic mechanisms as of comparable or equal importance to traditional determinants of success in pathogens (e.g AMR and virulence). The acquisition or changes in catabolic related genes clearly aids pathogenic states, however, I speculate that actually evolutionary drive of bacterial taxa to develop pathogenic characteristics may simply be to access the rich nutrient source of their hosts (Rohmer et al., 2014). The role of catabolism should be increasingly recognised as a priority equivalent to classical pathogenic factors and even into the future, due to their necessity to virulence states, could be used within anti-bacterial therapies (Rohmer et al., 2014). tGWAS has shone a light on the potential of catabolic related genes and changes within these to contribute to pathogen success. Coupled with the iron uptake genes, it is evident that the evolution of *S. sonnei* has streamlined the genome to exploit the human host resources to the fullest aiding its success. These intuitively beneficial factors highlight the validity of GWAS to be able to identify factors contributing to success even those which are under-investigated in the scientific community. tGWAS could help direct functional microbiology work to those factors most responsible for driving the success of naturally occurring pathogen populations.

5.3.3. *Why did I choose* S. flexneri *over* S. sonnei?

Although tGWAS was validated well, firstly with AMR and virulence determinants and then with genes with intuitively beneficial functions. Due to the time restrictions of this project only one of the tGWAS, either *S. sonnei* or *S. flexneri*, results would be able to be investigated in better detail due to the plethora of results generated. It was decided that although the

results within *S. sonnei* were of great interest, the *S. flexneri* dataset was to be pursued in more detail. The reason for this was the distribution of isolates across years. Specifically, 86% of *S. sonnei* Murray isolates were isolated from 1937 and so there was unintentional bias towards that specific year meaning the pre-antibiotic era in general was not well represented. I believe that this could mean some potential factors positively associated over time could have been missed during the tGWAS analyses. In addition to this the clonal replacement of *S. sonnei* potentially could potentially contribute as well as there would little opportunity to identify polyphyletic traits. *S. flexneri* on the other hand had 104% more isolates (n=45) which were not bias towards a singular year and therefore better represented the pre-antibiotic era. Furthermore, *S. flexneri* represents the major contributor to shigellosis in terms of case numbers so therefore is of greater public health interest. These results still represent exciting insight into factors contributing to the success of *S. sonnei* as a pathogen over time and these results should be considered for future investigation to fully get a 'helicopter view' of the success of *Shigella* species over time as well as for potentially informing on other enteric bacteria.

5.4. Insights gained into *S. flexneri*

S. flexneri represents the major global contributor to shigellosis and as such it is vital to understand what drives *S. flexneri* as a successful pathogen to aid its treatment and management. tGWAS of *S. flexneri* isolates over a 96-year time period revealed multiple intuitively beneficial genes for success of pathogens including the known factors AMR and virulence but also novel uncharacterised factors including a novel putative adhesin, adhesin Stv.

5.4.1. The good, the bad and the ugly of investigating the T3SS

tGWAS was highly validated for *S. flexneri* via the occurrence of multiple AMR and virulence determinants across all three GWAS feature types. Of particular note is how tGWAS highlighted the importance of the T3SS for *S. flexneri*. The T3SS is well documented as a crucial mechanism for *Shigella* invasion encoded on the pINV, mediating entry of the cytosol of epithelial cells (Bajunaid et al., 2020). Due to the documented importance of the T3SS and its observation throughout the three GWAS types, it was prudent to investigate the tGWAS results for the T3SS components. Clearly this work demonstrates its continued importance over time making it a vital mechanism to the success of *S. flexneri*. However, actual investigation of T3SS components came with challenges.

Although my attempts to characterise the effects of singular SNPs within ipaC and ipgD were unfruitful, SNPs can have a significant effect on the function of proteins e.g mutations in the QRDR locus cause heightened and novel resistance. There were several challenges I faced when trying to model these SNPs and characterise their effects on ipaC and ipgD. One of major challenges was the lack of solved crystal structures for ipaC and ipgD within the PDB. Within the PDB there are 37 protein structures deposited for *S. flexneri* related to the T3SS including structures for ipaD and mxiC, however, there are still many key T3SS components missing including ipaC and ipgD (Barta et al., 2017, Deane et al., 2008). The incomplete nature of all T3SS crystal structures and their undocumented conformational changes when bound with their targets makes it difficult to fully investigate some parts of the T3SS. I truly think a completed set of protein crystal structures for such an important system should be a focus for future research for *Shigella* but also due to the T3SS being imperative, relevant, and far-reaching in other bacteria.

With the lack of crystal structures available, it was necessary for me to produce predicted protein structures to continue my investigation. This in turn had its own related challenges.

The predicted protein modelling via Alphafold did yield predicted structures for these structures, however, there was low confidence in the ipaC structure and although ipgD yielded a better predicted structure, there were still portions with low to very low confidence. DALI searches comparing the predicted structures to the PDB database confirmed previous understanding that ipaC and ipgD were not present on the PDB. The highest matches for both proteins were to uncharacterised proteins with only 12% identity match and low z scores, highlighting the lack of similar proteins on the PDB so I was unable to confirm the validity of these predicted models. Despite the varying confidence levels of these predicted structures, I did choose to continue trying to characterise the SNPs and their effects on protein structure. However, I understood that any results I did receive would have the caveat that the structures may not be accurate.

Apart from the actual structure of the proteins, the last challenge was actually characterising the effect of the SNPs. All SNP characterisations were completed genotypically without experimental confirmation. There are multiple different bioinformatic tools to assess changes in stability and flexibility upon missense mutations within proteins such as DeepDDG, DynaMut2 and Site Directed Mutator (SDM) (Cao et al., 2019, Pandurangan et al., 2017, Rodrigues et al., 2021). DeepDDG's webserver predicts the stability change of protein point mutations utilising neural networks and was chosen due to its overall top position ranking when compared with other comparable servers (Cao et al., 2019). Calculated $\Delta\Delta G$ for ipaC and ipgD were -0.391 kcal/mol and -2.355 kcal/mol respectively. Anything between 0.5 to -0.5 is deemed negligible whilst < -0.5 is deemed destabilising and not advantageous at all for protein structures. The seemingly negligible contribution from the SNP within ipaC and the destabilising SNP within ipgD begs the question why these particular SNPs were so strongly positively associated with time. They do not seem to offer any advantage to *S. sonnei*. It is

possible that the potential inaccuracy of protein structures and/or the inability of solely genotypic predictions was not sufficient to fully characterise the extent of the effect of these SNPS on protein structures and their relative functions.

I do not believe that I have strong enough evidence to make any substantial claim about why these SNPs are important or their full characterisation within the T3SS. To further improve upon these investigations crystal structures for known virulence systems would be an immense asset coupled with the ability to experimentally characterise the SNPs. This, however, is a major undertaking and is easier said than done and would require complex experimental work. To fully solve a crystal structure the protein must be crystallized and through X-ray diffraction the structure of the protein is determined. Determination of 3D structures via X-ray crystallography is complex and sometimes can take 3 -5 years (Bhasin and Raghava, 2006). With over 20 parts of the T3SS as well as potential crystal structures needed to show conformational changes of proteins as they bind to their targets, this is a long-term project. In addition to this investigation and characterisation of the impact of specific SNPs require further experimental work such as circular dichroism spectroscopy and intrinsic fluorescence to explore conformation stability (Whitmore and Wallace, 2008). I still believe that SNPs hold great potential insight into the evolution of the T3SS, however, I believe that bioinformatic approaches may need laboratory validation to truly confirm impact on the functional capacity of T3SS apparatus.

5.4.2. The 'blackhole' of hypothetical genes

Of particular interest throughout the tGWAS analyses are the hypothetical genes which occur at equivalent statistical significance to AMR and virulence. These represent potential novel factors contributing to pathogen success. The exciting nature of this is that they could be novel targets for therapeutics and anti-bacterial therapies in the future. Understanding fully

the evolutionary arc of pathogens can truly help the treatment and management of global public health pathogens.

5.4.2.1. Exciting identification of adhesin Stv

One of the most promising and exciting results to come from tGWAS as a whole was the identification of the novel putative adhesin Stv observed on a small ~2.5 kb plasmid, pStv. Adhesin Stv represents a well substantiated important factor contributing to the success of *S. flexneri*. Since 1978, within *S. flexneri* PGs 1 and 6, adhesin Stv has been associated and stabilised throughout PG expansions, where acquisition was followed by clonal expansion, particularly for PG1. Furthermore, external validation of Stv as a contributor beyond *S. flexneri* was confirmed through investigation of *S. sonnei*. Stv was acquired and stabilised early (1943) in the emergence of the globally dominating Lineage 3, predating the TN7/Int2 cassette, although the potential implication of this will be discussed later. In the wider context, Stv was found in multiple different bacterial taxa including the wider *Enterobacteriaceae* species and more distantly related bacterial e.g *Bordetella* spp. In combination, the acquisition, stabilisation and extensive presence of Stv not only in *Shigella* but across other bacterial taxa provides plenty of evidentiary support for the importance and contribution of Stv to the success of *Shigella* as a global pathogen.

5.4.2.2. Experimental confirmation as the next step

Although genotypically there is well substantiated support for the novel putative function of Stv as an adhesin, there is currently no experimental confirmation of Stv's function. The assignment of the adhesin function was made based upon the stark similarities between functional domains of Stv and another *E. coli* protein (WP_205849698.1) also assigned a putative adhesin function. No published work related to the *E. coli* protein has been published and literature searches based on the tool PubServer and the protein sequence

similarly yielded no publications (Jaroszewski et al., 2014). The absence of publications confirms the lack of any experimental work to confirm the putative function of an adhesin for both Stv and WP_205849698.1. In general, however, structural and functional annotation based on the NCBI annotation pipeline uses protein family models, a hierarchical collection of evidence composed of Hidden Markov Model-based and BLAST-based protein families (HMMs and BlastRules) and Conserved Domain Database architectures (CDDs) (Tatusova et al., 2016, Haft et al., 2018, Li et al., 2021). This annotation pipeline has great success with the majority of predicted prokaryotic protein-coding genes are supported by homology to known proteins, as high as 96% for bacterial species such as *S. aureus* (Tatusova et al., 2016). Knowing the detailed database and supported homology I believe there is a good probability that experimental work would confirm the putative function. In addition to this Alex Bateman from the EMBL-EBI recently curated a preliminary protein family (PF21527 – soon to be released) which clustered adhesin Stv with filamentous haemagglutinin domains which are well known to be included domains observed within adhesins in *Bordetella* species, aiding attachment to epithelial cells of the respiratory tract (Locht et al., 1993). The preliminary protein family contains domains associated with adherence providing further support for its putative function. However, laboratory confirmation would be my suggested next step and is currently being worked on within the Bakery lab group. Molecular Koch's postulates are a set of experimental criteria that must be satisfied to show that a gene found within a pathogen encodes a product that contributes to disease (Falkow, 2004). In this instance deletion of the Stv gene within a selection of *S. sonnei* strains to generate a knock-out strain is currently underway followed by complementation with a plasmid. An adhesion assay, most likely with HeLa cells, will then be conducted to determine adhesion ability. The use of *S. sonnei* strains is due to the

fact that these do not contain the pINV and therefore allows sole measurement of adhesion without invasion.

5.4.2.3. A question of time

For *S. flexneri*, tGWAS has truly highlighted key factors positively associated with success over time. However, a major part of the reason I chose to develop tGWAS was due to the unique time span (approx. 100 years) that I was able to investigate due to the addition of the historical Murray Collection isolates. The first observation of adhesin Stv, however, was in 1978 within *S. flexneri*. This classes as the post-antibiotic era, after the historical isolate collection. This is an interesting observation and begs the question that if I had excluded the Murray Collection would Stv have still appeared within the factors positively associated with time. It is possible that the extended time-period with the Murray isolates facilitated the positive association by providing years of data where Stv was absent compared to from 1978 onwards where it became present resulting in a stronger association. To confirm this, however, tGWAS would need to completed excluding the historical collection. Comparisons between the two could elucidate the effects of having a longer time span. Regardless of whether Stv may been identified with the exclusion of the Murray Collection, I believe the historical isolates are invaluable to investigations of bacterial evolution over time.

5.4.2.4. A bright future

It is easy to perceive this field of study at a micro level but to gain a fuller perspective, a broader view must be taken. If the experimental work confirmed an adhesin function for Stv, it could potentially lead to Stv being incorporated into novel anti-bacterial therapies. In recent years there has been increasing interest and development of therapies targeting pathogens without the usage of antibiotics. One of those novel approaches is anti-adhesion therapies. The approach is conducted through the use of agents that interfere with the ability of the

bacteria to adhere to tissues of the host, since such adhesion is one of the initial stages of the infectious process (Ofek et al., 2003). There has been a wide variety of experiment work to validate this approach in multiple different animal models, validating the potential effectiveness of anti-adhesion therapies (Ofek et al., 2003, Asadi et al., 2019). The global nature of *Shigella* coupled with the increasing complication of AMR and the lack of a licensed vaccine, makes anti-adhesion therapy attractive as a treatment for this pathogen. Potential anti-adhesion therapies via varying mechanisms have been developed for closely related pathogens including *E. coli* and *Salmonella* species (Howell et al., 2010, Miladi et al., 2016, Fessele and Lindhorst, 2013). However, even with initial successes there has been a lack of extensive use. One considerable problem is the presence of multiple bacterial adhesion mechanisms and so blocking a singular mechanism may not fully inhibit the bacteria (Asadi et al., 2019). It is possible a combination of anti-adhesion molecules would be necessary. Other problems range from low affinity of free receptors to the presence of common epitopes within human hosts (Asadi et al., 2019). Despite these challenges, further research to develop broad specificity inhibitors or a combination of specific better designed agents could potentially act as a ground-breaking therapy to aid the global AMR problem for bacteria. Adhesin Stv could represent a potential target for anti-adhesion therapy not just for *Shigella* but for wider bacterial taxa.

5.5. Colicins: The new frontier for interbacterial competition?

Due to the prominent nature of small mobile genetic elements and intercellular competition factors found to be positively associated over time with the success of *Shigella* as a pathogen, it was prudent to further investigate such elements. Within *S. sonnei* Lineage 3, the advantageous *E. coli* killing ability is an epidemiologically important phenotype. This is highlighted by MSM Clade 3 which did not exhibit *E. coli* killing was outcompeted by *E. coli*

killing subclades (MSM Clades 1, 2, and 4) (Hawkey et al., 2021). The *E. coli* killing ability demonstrated by Lineage 3 *S. sonnei* was investigated due to the potential contribution of such small mobile genetic elements. The ruling out of the T6SS in contributing to this phenotype was a major discovery and highlighted the importance of other factors for interbacterial competition. Are colicins the new frontier in interbacterial competition?

5.5.1. Bioinformatics meets experimental work

Within Chapter 2 and 3 I did not get to do any functional work; however, I think the true success story of the research depicted in chapter 4 has to be the phenomenal amalgamation between bioinformatics and experimental work. Experimental work, the novel BPER methodology, first revealed the widely distributed *E. coli* killing phenotype within Lineage 3 *S. sonnei*. Bioinformatics was then utilised to identify factors contributing to this advantageous phenotype. GWAS analyses via two different approaches highlighted the potential contribution of colicins. Key colicin clusters were identified genotypically that were present within the killing isolates and absent within non-killing isolates and so potentially contributed to the killing phenotype. Other factors such as a T6SS were ruled out, further suggesting the importance of colicins for this ability. Coming full circle, experimental work, via mass spectrometry, confirmed that supernatants from all isolates (100%, 15 of 15) that displayed *E. coli* killing contained at least one contributing colicin in the total peptides identified in the samples and no colicin peptides were found from samples (n=2/2) with a non-killing phenotype. While there might be several mechanisms at play in the complex environments and microbial communities that the pathogens live in (Lindsay et al., 2015, Pop et al., 2014), my work demonstrated the crucial role colicins play in *S. sonnei* competition with *E. coli* at an epidemiological level. The amalgamation of bioinformatics and experimental work

identified, investigated and successfully found contributors to the *E. coli* phenotype, an ability beneficial to the success of *S. sonnei*.

5.5.2. The challenges

There were some definite challenges associated with investigating colicins at both a genotypically and at an experimental level. Genotypically, in order to screen all the isolates for the presence or absence of colicins, a collection of colicin sequences needed to be established. There is currently no published database for colicins and so collation of sequences was completed by myself by collating sequences from the ENA and from prominent papers related to colicins including Hahn-Löbmann et al. (2019). Although the database I collated held >10,000 colicin sequences from a wide range of bacterial taxa, the question remains – was my database fully representative of the full plethora of colicin and colicin related genes present in bacteria? For the purposes of this particular study where the aims were simply to identify potential general factors contributing to *E. coli* killing, I believe the database was more than sufficient. However, if a more detailed question was to be asked about the complete colicin profiles within *Shigella*, it may be prudent to do a more thorough collation of sequences.

One other challenge that was prominent within this investigation was the close similarity between different colicin sequences. Genotypically, this meant that tools utilised screening of isolates for colicins clustered them at 95% sequence identity producing approximately 145 clusters based on the 10,000 sequences. This meant that a more in-depth exploration was needed to further tease apart which specific colicins were present in each isolate, however, genotypically this was overcome and completed. Experimentally this posed more a challenge, especially during the mass spectrometry experiments completed by colleagues at the University of Strasbourg. The similarity of the protein sequences for different colicins

meant that there were shared peptides between colicins making identifying specific colicins during the investigation difficult. There were subset (all the identified peptides are shared by those proteins) and same set (some of the peptides are shared) colicin sequences where all or some of the identified peptides were shared between colicins so I was unable to say whether one or all of these colicins were present in the samples. For the purposes of this investigation where I simply wanted to demonstrate that colicins were being produced but not necessarily begin an in-depth analysis on any specific colicin, the mass spectrometry data was perfect and clearly experimentally demonstrated the production of these colicins by *S. sonnei*.

5.5.3. The T6SS conundrum

Within *S. sonnei* there has been a documented functional T6SS discovered (Anderson et al., 2017b). Due to the nature of the function of the T6SS, I had to rule out or quantify its contribution to the *E. coli* killing phenotype. I had access to a newly resequenced nanopore assembly of CIP106347 (the original strain the T6SS was described in) utilised during Anderson et al. (2017b) investigation into the *S. sonnei* T6SS. I encountered some difficulties in extracting the T6SS in the complete state shown within Anderson et al. (2017b). All core T6SS genes were detected, but there were some inconsistencies with the schematic in the original study. Specifically, there were frameshift mutations within several core T6SS genes (TssC, TssK, ClpV and TssM) resulting in premature stop codons. One possible explanation is that these are simply sequencing errors, however, the resequencing of CIP106347 resulted in a high quality sequence with good coverage. It is more plausible that these resulted due to sporadic pseudogenisation of disused loci. However, there was a transposon disruption of TssM as well as transposon disruptions of the putative effector genes downstream of each of the VgrGs. Just the transposon disruption of TssM alone should render the T6SS non-

functional. This was inconsistent with the reported functionality of the T6SS in Anderson et al. (2017b).

While it is possible these disruptions could be sequencing errors; however, this is highly unlikely and raises questions about the functionality of the T6SS. As evaluation of a functional T6SS would require *in vivo* work outside of the capacity of our laboratory, I elected to take an *in silico* approach of comparing the predicted core genes of the T6SS from the resequenced CIP106347 with the Lineage 3 *S. sonnei* isolates and comparing the genotypes of killing and non-killing isolates. One or more of the key components of the T6SS (i.e., transposase or N terminus of TssC) were absent in 100% of our isolates and CIP106347 suggesting that the T6SS in *S. sonnei* is unlikely to be functional. If all key components had been present the inconsistencies of the T6SS within CIP106347 would have posed a greater problem as I would have had to check for the functionality of the T6SS. The complete absence of one or more components confirms the unfunctional capacity of T6SS and so did not require examination of T6SS functionality.

5.5.4. The possibility of other factors

Although this investigation focused solely on the potential contribution of colicins there were obviously other significant hits which were observed during the GWAS analyses. These were not investigated due to the quantity of colicin genes observed highly within Pyseer and Scoary. It is possible that the prominent *E. coli* killing ability of *S. sonnei* Lineage 3 is due to a combination of different factors, of which colicins may play a role but are not solely responsible. There were other factors significantly associated with the phenotype including phage related genes and plasmid replication/transfer genes. It is possible that these genes aid replication and transfer of small genetic elements such as colicin plasmids or they have

other functions within *S. sonnei* or its various plasmids and so contribute to *E. coli* killing in another way.

The phage related genes within *S. sonnei* may play a role in bacterial virulence as well. It is well known that bacteriophages have complex relationships with their bacterial hosts and there is growing research to suggest that they do contribute to the fitness and virulence of their hosts (Schroven et al., 2020). Within *E. coli*, a closely related bacterial species to *Shigella*, phage's are responsible for the production of the highly toxic Shiga toxin (Fortier and Sekulovic, 2013). I was unable to determine the exact phage's that the genes belonged to but it is possible that the phage's are aiding bacterial virulence in some manner and so contributing to interbacterial competition.

5.6. Success of *Shigella* – A wider perspective

Throughout this thesis I have used novel and traditional GWAS approaches to identify a plethora of factors which contribute to the success of *Shigella* species as a global pathogen. The success of any bacterial species is a complex and interlaced web of factors which all play roles at various stages of successful infection. There is much more to the success of a pathogen, including *Shigella*, than the popularised published characteristics of AMR and virulence (Holt et al., 2012b, Hawkey et al., 2021, Connor et al., 2015). Although these contribute greatly to the success of pathogens throughout all pathogen states including stress, invasion, survival and replication I think it is wise to view the evolution of global public health pathogens at a broader perspective to identify innovative ways to aid treatment and pathogen management. Not only this but tGWAS as a novel approach could be applied to any bacterial taxa where isolates spanning years could be utilised.

Although AMR and virulence were known factors contributing to pathogen success over time the novel tGWAS approach revealed several novel insights. tGWAS highlighted the

importance of previously under researched factors contributing to success such as catabolism mechanism, iron metabolism, insertion sequences and the contribution of hypothetical genes. Adhesin Stv represents a novel gene associated with pathogenic success within *S. flexneri*, *S. sonnei* and other bacterial taxa. Its putative function as an adhesin highlights the importance of this mechanism for successful infection which is intuitively beneficial. In addition to this Stv highlights the potential importance of currently hypothetical genes which could be critical to long term success of public health pathogens. Furthermore, during this thesis novel insights into the contribution of colicins rather than that of the T6SS was revealed to be vital for the advantageous *E. coli* killing ability of *S. sonnei*, a key factor of interbacterial competition.

5.6.1. Species variation

Specifically, for *Shigella*, this work has demonstrated an abundance of known and unknown factors equally as important as AMR and virulence contributing to success. Interestingly, however, when you cross-compare the factors identified for both *S. flexneri* and *S. sonnei* over time there was very little overlap, if at all. The only real overlap was when both species identified genes associated with the citrate metabolic pathway, citA (*S. sonnei*) and citD (*S. flexneri*). Both of these genes were highly associated over time within their respective species, which puzzled me as *Shigella* species are typically considered citrate negative. However, with both species highlighting parts of this pathway, it must be contributing somewhat to *Shigella's* success over time. Citrate can act as an iron chelator contributing to iron transport systems as shown in *E. coli* (Mey et al., 2021). Since iron uptake was observed in tGWAS for both species, I hypothesise this may be a possible reason for the strong positive association over time within the two *Shigella* species.

With very limited overlap between the factors positively associated with *S. sonnei* and *S. flexneri*, this does then pose the question was I expecting more overlap? *S. flexneri* and *S. sonnei* are closely related genetically, however, they do dominate different geographical niches. *S. flexneri* is primarily isolated from LMIC, whereas *S. sonnei* is associated with economically developing and middle to high income countries. It is possible that the differing geographical niches, treatment regimens and environmental factors have led to slightly different evolutionary tracks. Varying selective pressures may have resulted in the different arcs of evolution resulting in the differing traits positively associated with time due to different niches and needs. Another possible factor which contributed to the small overlap between species was the clonality of *S. sonnei*. *S. sonnei* undergoes clonal replacement meaning that there is little opportunity to identify polyphyletic factors.

One way in which you could alter methodologies would be to combine the two species and conduct tGWAS. This may have revealed factors of success associated with both species; however, it would have likely missed species specific factors.

5.6.2. A broader insight into the evolutionary arc of AMR – lessons learnt from an adhesin

One of the major findings during this thesis was the identification of adhesin Stv. The prominent nature of Stv within both *S. flexneri* and *S. sonnei* as well as wider bacterial taxa provides substantial evidence for the importance of adhesion to the success of pathogens. Biologically this makes intuitive sense but one of the interesting points that occurred was within *S. sonnei*. Adhesin Stv was stabilised early in Lineage 3 prior to the AMR cassette, Tn7/Int2 cassette, thought to be the primary cause as to why Lineage 3 became the globally dominating Lineage (Holt et al., 2012b). I believe this could suggest that there could be precursors necessary for or to aid the acquisition of AMR determinants. This is an interesting

biological field of growing interest. Interrogations of pangenomes of the closely related *E. coli* identified significant co-occurrence hubs which were linked to virulence and mobile elements and between AMR and transposons (Hall et al., 2021). Furthermore, configurations of plasmid sequences highlighted how different optimal configurations allowed for better adaption to environments such as hospital settings exemplified by copper resistance, phosphotransferase systems, or bacteriocin genes potentially involved in niche adaptation (Alonso-del Valle et al., 2021, Arredondo-Alonso et al., 2020). This of similar thought to my hypothesis where Stv provided the opportunity for Stv-positive *S. sonnei* to further acquire mobilizable elements such as AMR and colicins, ultimately combining to drive the success of this global MDR lineage. I think that the evolution of AMR or other key adaptions to success are a precarious balance between many things including potential pre-requisites, fitness costs and selective pressures. I believe that these other factors are currently underestimated and under-investigated within the evolutionary genomics of AMR. This is an important area of work due to its potential as a stopping point / target for preventing AMR evolution by treatment and pathogen management strategies.

5.6.3. The importance of the 'blackhole' of hypothetical genes

I think the other major point when discussing the novel putative adhesin Stv is just how much bacterial genomes are shaped by presently unknown and uncharacterised genes such as adhesin Stv. The success of *Shigella* species is a multifaceted and complex web of factors which work in combination and separately to aid each part of infection. It was highlighted within *E. coli* pangenomes that hypothetical proteins have a large influence on the *E. coli* accessory genome, and I think my work here further supports that narrative (Hall et al., 2021). The quantity of hypothetical genes which occurred during tGWAS highlights the 'blackhole' of bacterial genes which clearly shape the success of pathogens. These represent possibly the

most exciting resource of novel exploration for the success of pathogens. These represent possible new targets for therapeutics not only for *Shigella* but for other closely related *Enterobacteriaceae* aiding the treatment and management of key pathogens.

The success of *Shigella* species has been successfully investigated during the tGWAS analyses as well as identifying potential factors contributing to the advantageous *E. coli* killing ability. I think though it is important to note that the factors identified here for *Shigella* species may also hold contributions for wider bacterial species including other *Enterobacteriaceae*, as seen with Stv.

5.7. Future work

For the sheer amount of data generated during this project there will clearly need to be further experimentation, research and interrogation to fully exploit the data. Bioinformatically, I believe the *S. sonnei* tGWAS dataset and the GWAS analysis for the *E. coli* killing ability could be further investigated in the same manner as *S. flexneri*. This could potentially highlight other key factors specific to this species and ability. Further to this I believe that of particular focus should be these so called 'blackhole' genes throughout both analyses as these represent possible novel factors completely unchartered in these species. In practice this may be difficult as bioinformatically, I had trouble identifying all unknown proteins and so potentially this may need to be combined with experimental work to get the best chance of success.

Experimentally there could be several key areas to target. Possibly the most important to this work, and which is currently ongoing, is to confirm the function of Stv as an adhesin. Knockout mutants are being currently being generated and an adhesion assay completed using HeLa cells to identify the adhesion capability of this protein. Another area to target experimentally would be through SNP investigations. As demonstrated through the challenges seen when

investigating the SNPs within ipaC and ipgD for *S. flexneri*, genotypic examinations alone may be insufficient to properly characterise and elucidate effect on protein structure and function. Experimental investigations such as circular dichroism spectroscopy and intrinsic fluorescence may help to substantiate characterisations. Coupled with this may be the need to truly define some key protein structures via solving their crystal structures. However, this a long and challenging process and so care should be taken to decide if this is a necessity, for example for the T3SS, or would the cost outweigh the benefit.

Supplementary Information and Tables

Links available to download data.

Supplementary Figure 1: Graphical representation of the decision trees allowing prioritisation of key GWAS feature types – SNPs, Kmers and COGs.

https://figshare.com/s/0e7562ab50f103a12ef0

Supplementary Figure 2: Confirmation of adhesin Stv on pStv by PCR. (A) A linear map of the contiguous sequence containing adehsin extracted from ERR1364216 is shown with gene annotations. Primer sequences designed to amplify the plasmid interior and exterior to the contig boundary as well as internal region of the adhesin gene are shown in coloured blocks, with the sequences and the expected product sizes from PCR in the table below. (B) Gel image of three PCR reactions along with their relevant no template controls for the respective primer pairs were conducted and demonstrated, through their being the expected size, that the sequence is complete. Based on the evidence from the PCR reactions and genome coverage information, overlapping contiguous sequences were used to manually edit a final circular pStv sequence, which is deposited in GenBank under accession number: OP113953
https://figshare.com/s/547dfa6e4fc844eb962c

Supplementary Figure 3: E. coli killing by representatives of S. sonnei MSM-associated clades. (A) Colony Forming Units on both non-selective (LB – top row) and selective (Kanamycin 30µg/ml – bottom row) for E. coli after the competition assay of individual representative isolates belonging to the different MSM associated clades showing the appearance of E. coli colonies when competed with only clade 3. (B) confirmation of the presence of E. coli colonies via the chromosomally encoded constitutively expressed GFP in E. coli MG1655 used in this study.

https://figshare.com/s/5c0fa01664e1fd531c23

Supplementary Figure 4: *E. coli* killing by the cell-free supernatants of selected representative *S. sonnei* isolates. *E. coli* killing as indicated by the zones of clearance of the *E. coli* lawn around the filter papers containing cell-free supernatant from selected *S. sonnei* isolates demonstrating *E. coli* killing via colicins present in the supernatants.

https://figshare.com/s/efb8965a0260d19b2e74

Supplementary Table 1: Details of the *S. sonnei* isolates used in this study including accession numbers, year, serotype, phylogroup and AMR and virulence profile.

https://figshare.com/s/075e2a679d633ab37589

Supplementary Table 2: Full GWAS analysis results for *S. sonnei*.

https://figshare.com/s/f1d0b33e656907a2183c

Supplementary Table 3: Details of the *S. flexneri* isolates used in this study including accession numbers, year, serotype, phylogroup and AMR and virulence profile.

https://figshare.com/s/29d478b20fba9eedce67

Supplementary Table 4: Full GWAS analysis results for *S. flexneri*.

https://figshare.com/s/289b8f324402f94632c5

Supplementary Table 5: Details of the *S. sonnei* Lineage 3 isolates used in this study including accession numbers, year, genotype, BPER results, Lineage and colicin profile.

https://figshare.com/s/d1a0eb9dcb7f70089b19

Supplementary Table 6: Full GWAS analysis results for *S. sonnei* Lineage 3 *E. coli* killing ability.

https://figshare.com/s/c511f4b3100aabedc0dd

Supplementary Table 7: Complete results of the mass spectrometry analysis of the *S. sonnei* representative isolates for the presence of colicins.

https://figshare.com/s/ae361f973c2586c4f40b

Bibliography

2019a. Picard toolkit. *Broad Institute, GitHub repository*.

2019b. Protein Data Bank: the single global archive for 3D macromolecular structure data. *Nucleic acids research,* 47**,** D520-D528.

AGHAPOUR, Z., GHOLIZADEH, P., GANBAROV, K., BIALVAEI, A. Z., MAHMOOD, S. S., TANOMAND, A., YOUSEFI, M., ASGHARZADEH, M., YOUSEFI, B. & KAFIL, H. S. 2019. Molecular mechanisms related to colistin resistance in Enterobacteriaceae. *Infect Drug Resist,* 12**,** 965-975.

AL-HASANI, K., HENDERSON, I. R., SAKELLARIS, H., RAJAKUMAR, K., GRANT, T., NATARO, J. P., ROBINS-BROWNE, R. & ADLER, B. 2000. The sigA gene which is borne on the she pathogenicity island of Shigella flexneri 2a encodes an exported cytopathic protease involved in intestinal fluid accumulation. *Infection and immunity,* 68**,** 2457-2463.

ALCOCK, B. P., RAPHENYA, A. R., LAU, T. T. Y., TSANG, K. K., BOUCHARD, M., EDALATMAND, A., HUYNH, W., NGUYEN, A. V., CHENG, A. A., LIU, S., MIN, S. Y., MIROSHNICHENKO, A., TRAN, H. K., WERFALLI, R. E., NASIR, J. A., OLONI, M., SPEICHER, D. J., FLORESCU, A., SINGH, B., FALTYN, M., HERNANDEZ-KOUTOUCHEVA, A., SHARMA, A. N., BORDELEAU, E., PAWLOWSKI, A. C., ZUBYK, H. L., DOOLEY, D., GRIFFITHS, E., MAGUIRE, F., WINSOR, G. L., BEIKO, R. G., BRINKMAN, F. S. L., HSIAO, W. W. L., DOMSELAAR, G. V. & MCARTHUR, A. G. 2020. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res,* 48**,** D517-d525.

ALDRED, K. J., KERNS, R. J. & OSHEROFF, N. 2014. Mechanism of quinolone action and resistance. *Biochemistry,* 53**,** 1565-1574.

ALLISON, G. E. & VERMA, N. K. 2000. Serotype-converting bacteriophages and O-antigen modification in Shigella flexneri. *Trends Microbiol,* 8**,** 17-23.

ALONSO-DEL VALLE, A., LEÓN-SAMPEDRO, R., RODRÍGUEZ-BELTRÁN, J., DELAFUENTE, J., HERNÁNDEZ-GARCÍA, M., RUIZ-GARBAJOSA, P., CANTÓN, R., PEÑA-MILLER, R. & SAN MILLÁN, A. 2021. Variability of plasmid fitness effects contributes to plasmid persistence in bacterial communities. *Nature Communications,* 12**,** 2653.

ALTSCHUL, S. F., GISH, W., MILLER, W., MYERS, E. W. & LIPMAN, D. J. 1990. Basic local alignment search tool. *Journal of Molecular Biology,* 215**,** 403-410.

AMARASINGHE, S. L., SU, S., DONG, X., ZAPPIA, L., RITCHIE, M. E. & GOUIL, Q. 2020. Opportunities and challenges in long-read sequencing data analysis. *Genome Biology,* 21**,** 30.

AMATO, S. M., FAZEN, C. H., HENRY, T. C., MOK, W. W., ORMAN, M. A., SANDVIK, E. L., VOLZING, K. G. & BRYNILDSEN, M. P. 2014. The role of metabolism in bacterial persistence. *Front Microbiol,* 5**,** 70.

AMBROSI, C., POMPILI, M., SCRIBANO, D., ZAGAGLIA, C., RIPA, S. & NICOLETTI, M. 2012. Outer membrane protein A (OmpA): a new player in shigella flexneri protrusion formation and inter-cellular spreading. *PLoS One,* 7**,** e49625.

ANDERSON, M., SANSONETTI, P. J. & MARTEYN, B. S. 2016. Shigella Diversity and Changing Landscape: Insights for the Twenty-First Century. *Front Cell Infect Microbiol,* 6**,** 45.

ANDERSON, M. C., VONAESCH, P., SAFFARIAN, A., MARTEYN, B. S. & SANSONETTI, P. J. 2017a. Shigella sonnei Encodes a Functional T6SS Used for Interbacterial Competition and Niche Occupancy. *Cell Host Microbe,* 21**,** 769-776 e3.

ANDERSON, M. C., VONAESCH, P., SAFFARIAN, A., MARTEYN, B. S. & SANSONETTI, P. J. 2017b. Shigella sonnei Encodes a Functional T6SS Used for Interbacterial Competition and Niche Occupancy. *Cell Host & Microbe,* 21**,** 769-776.e3.

ANDERSON, S. 1981. Shotgun DNA sequencing using cloned DNase I-generated fragments. *Nucleic acids research,* 9**,** 3015-3027.

ARREDONDO-ALONSO, S., TOP, J., MCNALLY, A., PURANEN, S., PESONEN, M., PENSAR, J., MARTTINEN, P., BRAAT, J. C., ROGERS, M. R. C., SCHAIK, W. V., KASKI, S., WILLEMS, R. J. L., CORANDER, J. & SCHÜRCH, A. C. 2020. Plasmids Shaped the Recent Emergence of the Major Nosocomial Pathogen Enterococcus faecium. *mBio,* 11**,** e03284-19.

ASADI, A., RAZAVI, S., TALEBI, M. & GHOLAMI, M. 2019. A review on anti-adhesion therapies of bacterial diseases. *Infection,* 47**,** 13-23.

AZMI, I. J., KHAJANCHI, B. K., AKTER, F., HASAN, T. N., SHAHNAIJ, M., AKTER, M., BANIK, A., SULTANA, H., HOSSAIN, M. A., AHMED, M. K., FARUQUE, S. M. & TALUKDER, K. A. 2014. Fluoroquinolone Resistance Mechanisms of Shigella flexneri Isolated in Bangladesh. *PLOS ONE,* 9**,** e102533.

BAJUNAID, W., HAIDAR-AHMAD, N., KOTTARAMPATEL, A. H., OURIDA MANIGAT, F., SILUÉ, N., TCHAGANG, C. F., TOMARO, K. & CAMPBELL-VALOIS, F. X. 2020. The T3SS of Shigella: Expression, Structure, Function, and Role in Vacuole Escape. *Microorganisms,* 8.

BAKER, K., DALLMAN, T., FIELD, N., CHILDS, T., MITCHELL, H., DAY, M., WEILL, F.-X., LEFÈVRE, S., TOURDJMAN, M., HUGHES, G., JENKINS, C. & THOMSON, N. 2018a. *Genomic epidemiology of Shigella in the United Kingdom shows transmission of pathogen sublineages and determinants of antimicrobial resistance*.

BAKER, K. S., BURNETT, E., MCGREGOR, H., DEHEER-GRAHAM, A., BOINETT, C., LANGRIDGE, G. C., WAILAN, A. M., CAIN, A. K., THOMSON, N. R., RUSSELL, J. E. & PARKHILL, J. 2015a. The Murray collection of pre-antibiotic era Enterobacteriacae: a unique research resource. *Genome Medicine,* 7**,** 97.

BAKER, K. S., CAMPOS, J., PICHEL, M., DELLA GASPERA, A., DUARTE-MARTINEZ, F., CAMPOS-CHACON, E., BOLANOS-ACUNA, H. M., GUZMAN-VERRI, C., MATHER, A. E., DIAZ VELASCO, S., ZAMUDIO ROJAS, M. L., FORBESTER, J. L., CONNOR, T. R., KEDDY, K. H., SMITH, A. M., LOPEZ DE DELGADO, E. A., ANGIOLILLO, G., CUAICAL, N., FERNANDEZ, J., AGUAYO, C., MORALES AGUILAR, M., VALENZUELA, C., MORALES MEDRANO, A. J., SIROK, A., WEILER GUSTAFSON, N., DIAZ GUEVARA, P. L., MONTANO, L. A., PEREZ, E. & THOMSON, N. R. 2017. Whole genome sequencing of Shigella sonnei through PulseNet Latin America and Caribbean: advancing global surveillance of foodborne illnesses. *Clin Microbiol Infect,* 23**,** 845-853.

BAKER, K. S., DALLMAN, T. J., ASHTON, P. M., DAY, M., HUGHES, G., CROOK, P. D., GILBART, V. L., ZITTERMANN, S., ALLEN, V. G. & HOWDEN, B. P. 2015b. Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. *The Lancet infectious diseases,* 15**,** 913-921.

BAKER, K. S., DALLMAN, T. J., ASHTON, P. M., DAY, M., HUGHES, G., CROOK, P. D., GILBART, V. L., ZITTERMANN, S., ALLEN, V. G., HOWDEN, B. P., TOMITA, T., VALCANIS, M., HARRIS, S. R., CONNOR, T. R., SINTCHENKO, V., HOWARD, P., BROWN, J. D., PETTY, N. K., GOUALI, M., THANH, D. P., KEDDY, K. H., SMITH, A. M., TALUKDER, K. A., FARUQUE, S. M., PARKHILL, J., BAKER, S., WEILL, F.-X., JENKINS, C. & THOMSON, N. R. 2015c. Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. *The Lancet Infectious Diseases,* 15**,** 913-921.

BAKER, K. S., DALLMAN, T. J., BEHAR, A., WEILL, F.-X., GOUALI, M., SOBEL, J., FOOKES, M., VALINSKY, L., GAL-MOR, O., CONNOR, T. R., NISSAN, I., BERTRAND, S., PARKHILL, J., JENKINS, C., COHEN, D. & THOMSON, N. R. 2016. Travel- and Community-Based Transmission of Multidrug-Resistant Shigella sonnei Lineage among International Orthodox Jewish Communities. 22**,** 1545-1553.

BAKER, K. S., DALLMAN, T. J., FIELD, N., CHILDS, T., MITCHELL, H., DAY, M., WEILL, F.-X., LEFÈVRE, S., TOURDJMAN, M., HUGHES, G., JENKINS, C. & THOMSON, N. 2018b. Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species. *Nature Communications,* 9**,** 1462.

BAKER, K. S., DALLMAN, T. J., FIELD, N., CHILDS, T., MITCHELL, H., DAY, M., WEILL, F. X., LEFEVRE, S., TOURDJMAN, M., HUGHES, G., JENKINS, C. & THOMSON, N. 2018c. Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species. *Nat Commun,* 9**,** 1462.

BAKER, K. S., MATHER, A. E., MCGREGOR, H., COUPLAND, P., LANGRIDGE, G. C., DAY, M., DEHEER-GRAHAM, A., PARKHILL, J., RUSSELL, J. E. & THOMSON, N. R. 2014. The extant World War 1 dysentery bacillus NCTC1: a genomic analysis. *The Lancet,* 384**,** 1691-1697.

BALAKRISHNAN, V. S. 2022. WHO's global genomic surveillance strategy. *The Lancet Infectious Diseases,* 22**,** 772.

BANKEVICH, A., NURK, S., ANTIPOV, D., GUREVICH, A. A., DVORKIN, M., KULIKOV, A. S., LESIN, V. M., NIKOLENKO, S. I., PHAM, S., PRJIBELSKI, A. D., PYSHKIN, A. V., SIROTKIN, A. V., VYAHHI, N., TESLER, G., ALEKSEYEV, M. A. & PEVZNER, P. A. 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,* 19**,** 455-477.

BARDEY, V., VALLET, C., ROBAS, N., CHARPENTIER, B., THOUVENOT, B., MOUGIN, A., HAJNSDORF, E., RÉGNIER, P., SPRINGER, M. & BRANLANT, C. 2005. Characterization of the molecular mechanisms involved in the differential production of erythrose-4-phosphate dehydrogenase, 3-phosphoglycerate

kinase and class II fructose-1,6-bisphosphate aldolase in Escherichia coli. *Molecular Microbiology,* 57**,** 1265-1287.
BARDHAN, P., FARUQUE, A. S., NAHEED, A. & SACK, D. A. 2010. Decrease in shigellosis-related deaths without Shigella spp.-specific interventions, Asia. *Emerg Infect Dis,* 16**,** 1718-23.
BARDSLEY, M., JENKINS, C., MITCHELL, H. D., MIKHAIL, A. F. W., BAKER, K. S., FOSTER, K., HUGHES, G. & DALLMAN, T. J. 2020. Persistent Transmission of Shigellosis in England Is Associated with a Recently Emerged Multidrug-Resistant Strain of Shigella sonnei. *J Clin Microbiol,* 58.
BARLOW, M. & HALL, B. G. J. J. O. M. E. 2002. Phylogenetic Analysis Shows That the OXA b-Lactamase Genes Have Been on Plasmids for Millions of Years. 55**,** 314-321.
BARTA, M. L., SHEARER, J. P., ARIZMENDI, O., TREMBLAY, J. M., MEHZABEEN, N., ZHENG, Q., BATTAILE, K. P., LOVELL, S., TZIPORI, S., PICKING, W. D., SHOEMAKER, C. B. & PICKING, W. L. 2017. Single-domain antibodies pinpoint potential targets within Shigella invasion plasmid antigen D of the needle tip complex for inhibition of type III secretion. *J Biol Chem,* 292**,** 16677-16687.
BELTRAMETTI, F., KRESSE, A. U. & GUZMÁN, C. A. 1999. Transcriptional regulation of the esp genes of enterohemorrhagic Escherichia coli. *J Bacteriol,* 181**,** 3409-18.
BENGTSSON, R. J., SIMPKIN, A. J., PULFORD, C. V., LOW, R., RASKO, D. A., RIGDEN, D. J., HALL, N., BARRY, E. M., TENNANT, S. M. & BAKER, K. S. 2022. Pathogenomic analyses of Shigella isolates inform factors limiting shigellosis prevention and control across LMICs. *Nature Microbiology,* 7**,** 251-261.
BENNETT, R. J. & BAKER, K. S. 2019. Looking backwards to move forward: the utility of sequencing historical bacterial genomes. *J Clin Microbiol*.
BENNETT, R. J., DE SILVA, P. M., BENGTSSON, R. J., HORSBURGH, M. J., BLOWER, T. R. & BAKER, K. S. 2022. Temporal GWAS identifies a widely distributed putative adhesin contributing to pathogen success in Shigella spp. *bioRxiv***,** 2022.08.23.504947.
BERMAN, H. M., WESTBROOK, J., FENG, Z., GILLILAND, G., BHAT, T. N., WEISSIG, H., SHINDYALOV, I. N. & BOURNE, P. E. 2000. The Protein Data Bank. *Nucleic Acids Research,* 28**,** 235-242.
BERNARDINI, M. L., FONTAINE, A. & SANSONETTI, P. J. 1990. The two-component regulatory system ompR-envZ controls the virulence of Shigella flexneri. *J Bacteriol,* 172**,** 6274-81.
BHASIN, M. & RAGHAVA, G. P. S. 2006. 8 - Computational Methods in Genome Research. *In:* ARORA, D. K., BERKA, R. M. & SINGH, G. B. (eds.) *Applied Mycology and Biotechnology.* Elsevier.
BHENDE, P. M. & EGAN, S. M. 2000. Genetic evidence that transcription activation by RhaS involves specific amino acid contacts with sigma 70. *J Bacteriol,* 182**,** 4959-69.
BHUNIA, A. K. 2018. Shigella Species. *Foodborne Microbial Pathogens: Mechanisms and Pathogenesis.* New York, NY: Springer New York.

BIRCH, L. C. 1957. The Meanings of Competition. *The American Naturalist,* 91**,** 5-18.
BISHOP, J. G., 3RD & HUNT, J. A. 1988. DNA divergence in and around the alcohol dehydrogenase locus in five closely related species of Hawaiian Drosophila. *Mol Biol Evol,* 5**,** 415-31.
BOLGER, A. M., LOHSE, M. & USADEL, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.
BORG, M. L., MODI, A., TOSTMANN, A., GOBIN, M., CARTWRIGHT, J., QUIGLEY, C., CROOK, P., BOXALL, N., PAUL, J., CHEASTY, T., GILL, N., HUGHES, G., SIMMS, I. & OLIVER, I. 2012. Ongoing outbreak of Shigella flexneri serotype 3a in men who have sex with men in England and Wales, data from 2009-2011. *Euro Surveill,* 17.
BORTOLAIA, V., KAAS, R. S., RUPPE, E., ROBERTS, M. C., SCHWARZ, S., CATTOIR, V., PHILIPPON, A., ALLESOE, R. L., REBELO, A. R., FLORENSA, A. F., FAGELHAUER, L., CHAKRABORTY, T., NEUMANN, B., WERNER, G., BENDER, J. K., STINGL, K., NGUYEN, M., COPPENS, J., XAVIER, B. B., MALHOTRA-KUMAR, S., WESTH, H., PINHOLT, M., ANJUM, M. F., DUGGETT, N. A., KEMPF, I., NYKÄSENOJA, S., OLKKOLA, S., WIECZOREK, K., AMARO, A., CLEMENTE, L., MOSSONG, J., LOSCH, S., RAGIMBEAU, C., LUND, O. & AARESTRUP, F. M. 2020. ResFinder 4.0 for predictions of phenotypes from genotypes. *Journal of Antimicrobial Chemotherapy,* 75**,** 3491-3500.
BOUMGHAR-BOURTCHAI, L., MARIANI-KURKDJIAN, P., BINGEN, E., FILLIOL, I., DHALLUIN, A., IFRANE, S. A., WEILL, F.-X. & LECLERCQ, R. 2008. Macrolide-resistant Shigella sonnei. *Emerging infectious diseases,* 14**,** 1297-1299.
BRAVO, V., PUHAR, A., SANSONETTI, P., PARSOT, C. & TORO, C. S. 2015. Distinct mutations led to inactivation of type 1 fimbriae expression in Shigella spp. *PLoS One,* 10**,** e0121785.
BRENNER, D. J., KRIEG, N. R., STALEY, J. T. & GARRITY, G. 2005. *Bergey's Manual® of Systematic Bacteriology: Volume Two The Proteobacteria Part C The Alpha-, Beta-, Delta-, and Epsilonproteobacteria*, Springer.
BRYNILDSRUD, O., BOHLIN, J., SCHEFFER, L. & ELDHOLM, V. 2016a. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol,* 17**,** 238.
BRYNILDSRUD, O., BOHLIN, J., SCHEFFER, L. & ELDHOLM, V. 2016b. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biology,* 17**,** 238.
BRZÓSTKOWSKA, M., RACZKOWSKA, A. & BRZOSTEK, K. 2012. OmpR, a response regulator of the two-component signal transduction pathway, influences inv gene expression in Yersinia enterocolitica O9. *Front Cell Infect Microbiol,* 2**,** 153.
BUNIELLO, A., MACARTHUR, J. A L., CEREZO, M., HARRIS, L. W., HAYHURST, J., MALANGONE, C., MCMAHON, A., MORALES, J., MOUNTJOY, E., SOLLIS, E., SUVEGES, D., VROUSGOU, O., WHETZEL, P. L., AMODE, R., GUILLEN, J. A., RIAT, H. S., TREVANION, S. J., HALL, P., JUNKINS, H., FLICEK, P., BURDETT, T.,

HINDORFF, L. A., CUNNINGHAM, F. & PARKINSON, H. 2018. The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Research,* 47**,** D1005-D1012.

BURKINSHAW, B. J., PREHNA, G., WORRALL, L. J. & STRYNADKA, N. C. 2012. Structure of Salmonella effector protein SopB N-terminal domain in complex with host Rho GTPase Cdc42. *J Biol Chem,* 287**,** 13348-55.

CABONI, M., PÉDRON, T., ROSSI, O., GOULDING, D., PICKARD, D., CITIULO, F., MACLENNAN, C. A., DOUGAN, G., THOMSON, N. R., SAUL, A., SANSONETTI, P. J. & GERKE, C. 2015. An O antigen capsule modulates bacterial pathogenesis in Shigella sonnei. *PLoS Pathog,* 11**,** e1004749.

CAMPBELL, J. W., MORGAN-KISS, R. M. & E. CRONAN JR, J. 2003. A new Escherichia coli metabolic competency: growth on fatty acids by a novel anaerobic β-oxidation pathway. *Molecular Microbiology,* 47**,** 793-805.

CAO, H., WANG, J., HE, L., QI, Y. & ZHANG, J. Z. 2019. DeepDDG: Predicting the Stability Change of Protein Point Mutations Using Neural Networks. *Journal of Chemical Information and Modeling,* 59**,** 1508-1514.

CARATTOLI, A. 2009. Resistance plasmid families in Enterobacteriaceae. *Antimicrobial agents and chemotherapy,* 53**,** 2227-2238.

CARAYOL, N. & TRAN VAN NHIEU, G. 2013. The inside story of Shigella invasion of intestinal epithelial cells. *Cold Spring Harbor perspectives in medicine,* 3**,** a016717-a016717.

CARBONETTI, N. H. & WILLIAMS, P. H. 1984. A cluster of five genes specifying the aerobactin iron uptake system of plasmid ColV-K30. *Infect Immun,* 46**,** 7-12.

CARNEIRO, L. A., TRAVASSOS, L. H., SOARES, F., TATTOLI, I., MAGALHAES, J. G., BOZZA, M. T., PLOTKOWSKI, M. C., SANSONETTI, P. J., MOLKENTIN, J. D. & PHILPOTT, D. J. 2009. Shigella induces mitochondrial dysfunction and cell death in nonmyleoid cells. *Cell host & microbe,* 5**,** 123-136.

CASALINO, M., PROSSEDA, G., BARBAGALLO, M., IACOBINO, A., CECCARINI, P., LATELLA, M. C., NICOLETTI, M. & COLONNA, B. 2010. Interference of the CadC regulator in the arginine-dependent acid resistance system of Shigella and enteroinvasive E. coli. *Int J Med Microbiol,* 300**,** 289-95.

CASCALES, E., BUCHANAN, S. K., DUCHÉ, D., KLEANTHOUS, C., LLOUBÈS, R., POSTLE, K., RILEY, M., SLATIN, S. & CAVARD, D. 2007. Colicin biology. *Microbiol Mol Biol Rev,* 71**,** 158-229.

CAZA, M. & KRONSTAD, J. W. 2013. Shared and distinct mechanisms of iron acquisition by bacterial and fungal pathogens of humans. *Front Cell Infect Microbiol,* 3**,** 80.

CESARENI, G., MUESING, M. A. & POLISKY, B. 1982. Control of ColE1 DNA replication: the rop gene product negatively affects transcription from the replication primer promoter. *Proc Natl Acad Sci U S A,* 79**,** 6313-7.

CHANG, Z., LU, S., CHEN, L., JIN, Q. & YANG, J. 2012. Causative Species and Serotypes of Shigellosis in Mainland China: Systematic Review and Meta-Analysis. *PLOS ONE,* 7**,** e52515.

CHANIN, R. B., NICKERSON, K. P., LLANOS-CHEA, A., SISTRUNK, J. R., RASKO, D. A., KUMAR, D. K. V., DE LA PARRA, J., AUCLAIR, J. R., DING, J., LI, K., DOGIPARTHI, S. K., KUSBER, B. J. D. & FAHERTY, C. S. 2019a. Shigella flexneri Adherence Factor Expression in In Vivo-Like Conditions. *mSphere,* 4.

CHANIN, R. B., NICKERSON, K. P., LLANOS-CHEA, A., SISTRUNK, J. R., RASKO, D. A., KUMAR, D. K. V., PARRA, J. D. L., AUCLAIR, J. R., DING, J., LI, K., DOGIPARTHI, S. K., KUSBER, B. J. D. & FAHERTY, C. S. 2019b. Shigella flexneri adherence factor expression in in vivo-like conditions. *bioRxiv*, 514679.

CHARLES, H., PROCHAZKA, M., THORLEY, K., CREWDSON, A., GREIG, D. R., JENKINS, C., PAINSET, A., FIFER, H., BROWNING, L., CABREY, P., SMITH, R., RICHARDSON, D., WATERS, L., SINKA, K., GODBOLE, G. & OUTBREAK CONTROL, T. 2022. Outbreak of sexually transmitted, extensively drug-resistant Shigella sonnei in the UK, 2021-22: a descriptive epidemiological study. *Lancet Infect Dis*.

CHEN, L., ZHENG, D., LIU, B., YANG, J. & JIN, Q. 2015. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Research,* 44, D694-D697.

CHIOU, C. S., IZUMIYA, H., KAWAMURA, M., LIAO, Y. S., SU, Y. S., WU, H. H., CHEN, W. C. & LO, Y. C. 2016. The worldwide spread of ciprofloxacin-resistant Shigella sonnei among HIV-infected men who have sex with men, Taiwan. *Clinical Microbiology and Infection,* 22, 383.e11-383.e16.

CHOI, S. Y., JEON, Y. S., LEE, J. H., CHOI, B., MOON, S. H., VON SEIDLEIN, L., CLEMENS, J. D., DOUGAN, G., WAIN, J., YU, J., LEE, J. C., SEOL, S. Y., LEE, B. K., SONG, J. H., SONG, M., CZERKINSKY, C., CHUN, J. & KIM, D. W. 2007. Multilocus sequence typing analysis of Shigella flexneri isolates collected in Asian countries. *J Med Microbiol,* 56, 1460-1466.

CHUNG THE, H., BOINETT, C., PHAM THANH, D., JENKINS, C., WEILL, F.-X., HOWDEN, B. P., VALCANIS, M., DE LAPPE, N., CORMICAN, M. & WANGCHUK, S. 2019. Dissecting the molecular evolution of fluoroquinolone-resistant Shigella sonnei. *Nature communications,* 10, 1-13.

CHUNG THE, H., RABAA, M. A., PHAM THANH, D., DE LAPPE, N., CORMICAN, M., VALCANIS, M., HOWDEN, B. P., WANGCHUK, S., BODHIDATTA, L., MASON, C. J., NGUYEN THI NGUYEN, T., VU THUY, D., THOMPSON, C. N., PHU HUONG LAN, N., VOONG VINH, P., HA THANH, T., TURNER, P., SAR, P., THWAITES, G., THOMSON, N. R., HOLT, K. E. & BAKER, S. 2016. South Asia as a Reservoir for the Global Spread of Ciprofloxacin-Resistant Shigella sonnei: A Cross-Sectional Study. *PLoS medicine,* 13, e1002055-e1002055.

CHUNG THE, H., RABAA, M. A., THANH, D. P., RUEKIT, S., WANGCHUK, S., DORJI, T., TSHERING, K. P., NGUYEN, T. N. T., VINH, P. V., THANH, T. H., MINH, C. N. N., TURNER, P., SAR, P., THWAITES, G., HOLT, K. E., THOMSON, N. R., BODHIDATTA, L., JEFFRIES MASON, C. & BAKER, S. 2015. Introduction and establishment of fluoroquinolone-resistant Shigella sonnei into Bhutan. *Microbial genomics,* 1, e000042-e000042.

CIANFANELLI, F. R., MONLEZUN, L. & COULTHURST, S. J. 2016. Aim, Load, Fire: The Type VI Secretion System, a Bacterial Nanoweapon. *Trends Microbiol,* 24**,** 51-62.

CINGOLANI, P., PLATTS, A., WANG LE, L., COON, M., NGUYEN, T., WANG, L., LAND, S. J., LU, X. & RUDEN, D. M. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly (Austin),* 6**,** 80-92.

COBURN, B., SEKIROV, I. & FINLAY, B. B. 2007. Type III secretion systems and disease. *Clin Microbiol Rev,* 20**,** 535-49.

COHEN, D., BASSAL, R., GOREN, S., ROUACH, T., TARAN, D., SCHEMBERG, B., PELED, N., KENESS, Y., KEN-DROR, S., VASILEV, V., NISSAN, I., AGMON, V. & SHOHAT, T. 2014. Recent trends in the epidemiology of shigellosis in Israel. *Epidemiol Infect,* 142**,** 2583-94.

COLL, F., MCNERNEY, R., GUERRA-ASSUNÇÃO, J. A., GLYNN, J. R., PERDIGÃO, J., VIVEIROS, M., PORTUGAL, I., PAIN, A., MARTIN, N. & CLARK, T. G. 2014. A robust SNP barcode for typing Mycobacterium tuberculosis complex strains. *Nature communications,* 5**,** 1-5.

COLLINS, C. & DIDELOT, X. 2018. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination. *PLOS Computational Biology,* 14**,** e1005958.

CONNOR, T. R., BARKER, C. R., BAKER, K. S., WEILL, F.-X., TALUKDER, K. A., SMITH, A. M., BAKER, S., GOUALI, M., THANH, D. P. & AZMI, I. J. J. E. 2015. Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri. 4**,** e07335.

CONOVER, M. S., HADJIFRANGISKOU, M., PALERMO, J. J., HIBBING, M. E., DODSON, K. W. & HULTGREN, S. J. 2016. Metabolic Requirements of Escherichia coli in Intracellular Bacterial Communities during Urinary Tract Infection Pathogenesis. *mBio,* 7**,** e00104-16.

COULTHURST, S. 2019. The Type VI secretion system: a versatile bacterial weapon. *Microbiology (Reading),* 165**,** 503-515.

CROUCHER, N. J., PAGE, A. J., CONNOR, T. R., DELANEY, A. J., KEANE, J. A., BENTLEY, S. D., PARKHILL, J. & HARRIS, S. R. 2014. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research,* 43**,** e15-e15.

DAHLBERG, C. & CHAO, L. 2003. Amelioration of the cost of conjugative plasmid carriage in Eschericha coli K12. *Genetics,* 165**,** 1641-1649.

DALDAL, F. 1984. Nucleotide sequence of gene pfkB encoding the minor phosphofructokinase of Escherichia coli K-12. *Gene,* 28**,** 337-42.

DALLMAN, T. J., CHATTAWAY, M. A., MOOK, P., GODBOLE, G., CROOK, P. D. & JENKINS, C. 2016. Use of whole-genome sequencing for the public health surveillance of Shigella sonnei in England and Wales, 2015. *J Med Microbiol,* 65**,** 882-4.

DANECEK, P. & MCCARTHY, S. A. 2017. BCFtools/csq: haplotype-aware variant consequences. *Bioinformatics,* 33**,** 2037-2039.

DAS, A. & MANDAL, J. 2019. Extensive inter-strain diversity among clinical isolates of Shigella flexneri with reference to its serotype, virulence traits and plasmid incompatibility types, a study from south India over a 6-year period. *Gut Pathogens,* 11**,** 33.

DATTA, N. & HUGHES, V. M. 1983. Plasmids of the same Inc groups in Enterobacteria before and after the medical use of antibiotics. *Nature,* 306**,** 616-7.

DAVISON, W. C. 1922. A bacteriological and clinical consideration of bacillary dysentery in adults and children. *Medicine,* 1**,** 389-510.

DE SCHRIJVER, K., BERTRAND, S., GUTIERREZ GARITANO, I., VAN DEN BRANDEN, D. & VAN SCHAEREN, J. 2011. Outbreak of Shigella sonnei infections in the Orthodox Jewish community of Antwerp, Belgium, April to August 2008. *Euro Surveill,* 16.

DEANE, J. E., ROVERSI, P., KING, C., JOHNSON, S. & LEA, S. M. 2008. Structures of the Shigella flexneri type 3 secretion system protein MxiC reveal conformational variability amongst homologues. *J Mol Biol,* 377**,** 985-92.

DEEN, J., MENGEL, M. A. & CLEMENS, J. D. 2020. Epidemiology of cholera. *Vaccine,* 38**,** A31-A40.

DIENEMANN, C., BØGGILD, A., WINTHER, K. S., GERDES, K. & BRODERSEN, D. E. 2011. Crystal structure of the VapBC toxin–antitoxin complex from Shigella flexneri reveals a hetero-octameric DNA-binding assembly. *Journal of molecular biology,* 414**,** 713-722.

DRANCOURT, M., ABOUDHARAM, G., SIGNOLI, M., DUTOUR, O. & RAOULT, D. 1998. Detection of 400-year-old Yersinia pestis DNA in human dental pulp: an approach to the diagnosis of ancient septicemia. *Proceedings of the National Academy of Sciences of the United States of America,* 95**,** 12637-12640.

DYK, T. K. V., TEMPLETON, L. J., CANTERA, K. A., SHARPE, P. L. & SARIASLANI, F. S. 2004. Characterization of the Escherichia coli AaeAB Efflux Pump: a Metabolic Relief Valve? *Journal of Bacteriology,* 186**,** 7196-7204.

EDGAR, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res,* 32**,** 1792-7.

ELIZABETH, R., BAISHYA, S., KALITA, B., WANGKHEIMAYUM, J., CHOUDHURY, M. D., CHANDA, D. D. & BHATTACHARJEE, A. 2022. Colistin exposure enhances expression of eptB in colistin-resistant Escherichia coli co-harboring mcr-1. *Scientific Reports,* 12**,** 1348.

ERIKSSON, S., LUCCHINI, S., THOMPSON, A., RHEN, M. & HINTON, J. C. 2003. Unravelling the biology of macrophage infection by gene expression profiling of intracellular Salmonella enterica. *Molecular microbiology,* 47**,** 103-118.

ESSA, A. M. M., JULIAN, D. J., KIDD, S. P., BROWN, N. L. & HOBMAN, J. L. 2003. Mercury Resistance Determinants Related to Tn21, Tn1696, and Tn5053 in Enterobacteria from the Preantibiotic Era. 47**,** 1115-1119.

ESTIMATES, G. H. 2016. Disease burden by Cause, Age, Sex, by Country and by Region, 2000-2015. *Geneva: World Health Organization*.

ESWARAMOORTHY, S., POULAIN, S., HIENERWADEL, R., BREMOND, N., SYLVESTER, M. D., ZHANG, Y.-B., BERTHOMIEU, C., VAN DER LELIE, D. & MATIN, A. 2012. Crystal Structure of ChrR—A Quinone Reductase with the Capacity to Reduce Chromate. *PLOS ONE,* 7**,** e36017.

FAHERTY, C., HARPER, J. M., SHEA-DONOHUE, T., BARRY, E. M., KAPER, J. B., FASANO, A. & NATARO, J. P. 2012. Chromosomal and Plasmid-Encoded Factors of Shigella flexneri Induce Secretogenic Activity Ex Vivo. *PLOS ONE,* 7**,** e49980.

FAHERTY, C. S., MERRELL, D. S., SEMINO-MORA, C., DUBOIS, A., RAMASWAMY, A. V. & MAURELLI, A. T. 2010. Microarray analysis of Shigella flexneri-infected epithelial cells identifies host factors important for apoptosis inhibition. *BMC genomics,* 11**,** 272-272.

FALKOW, S. 2004. Molecular Koch's postulates applied to bacterial pathogenicity — a personal recollection 15 years later. *Nature Reviews Microbiology,* 2**,** 67-72.

FALUSH, D. & BOWDEN, R. 2006. Genome-wide association mapping in bacteria? *Trends in Microbiology,* 14**,** 353-355.

FARHAT, M. R., FRESCHI, L., CALDERON, R., IOERGER, T., SNYDER, M., MEEHAN, C. J., DE JONG, B., RIGOUTS, L., SLOUTSKY, A., KAUR, D., SUNYAEV, S., VAN SOOLINGEN, D., SHENDURE, J., SACCHETTINI, J. & MURRAY, M. 2019. GWAS for quantitative resistance phenotypes in Mycobacterium tuberculosis reveals resistance genes and regulatory regions. *Nature Communications,* 10**,** 2128.

FELDGARDEN, M., BROVER, V., HAFT, D. H., PRASAD, A. B., SLOTTA, D. J., TOLSTOY, I., TYSON, G. H., ZHAO, S., HSU, C.-H., MCDERMOTT, P. F., TADESSE, D. A., MORALES, C., SIMMONS, M., TILLMAN, G., WASILENKO, J., FOLSTER, J. P. & KLIMKE, W. 2019. Validating the AMRFinder Tool and Resistance Gene Database by Using Antimicrobial Resistance Genotype-Phenotype Correlations in a Collection of Isolates. *Antimicrobial agents and chemotherapy,* 63**,** e00483-19.

FENG, Y. & CRONAN, J. E. 2014. PdhR, the pyruvate dehydrogenase repressor, does not regulate lipoic acid synthesis. *Res Microbiol,* 165**,** 429-38.

FESSELE, C. & LINDHORST, T. K. 2013. Effect of aminophenyl and aminothiahexyl α-d-glycosides of the manno-, gluco-, and galacto-series on type 1 fimbriae-mediated adhesion of Escherichia coli. *Biology,* 2**,** 1135-1149.

FINLAY, B. B. & MCFADDEN, G. 2006. Anti-Immunology: Evasion of the Host Immune System by Bacterial and Viral Pathogens. *Cell,* 124**,** 767-782.

FISCHER, N., MAEX, M., MATTHEUS, W., VAN DEN BOSSCHE, A., VAN CAUTEREN, D., LAISNEZ, V., HAMMAMI, N. & CEYSSENS, P. J. 2021. Genomic epidemiology of persistently circulating MDR Shigella sonnei strains associated with men who have sex with men (MSM) in Belgium (2013-19). *J Antimicrob Chemother,* 77**,** 89-97.

FISHER, C. R., DAVIES, N. M., WYCKOFF, E. E., FENG, Z., OAKS, E. V. & PAYNE, S. M. 2009. Genetics and virulence association of the Shigella flexneri sit iron transport system. *Infect Immun,* 77**,** 1992-9.

FLAJNIK, M. F. & KASAHARA, M. 2010. Origin and evolution of the adaptive immune system: genetic events and selective pressures. *Nature Reviews Genetics,* 11**,** 47-59.

FLEISCHMANN, R., ADAMS, M., WHITE, O., CLAYTON, R., KIRKNESS, E., KERLAVAGE, A., BULT, C., TOMB, J., DOUGHERTY, B., MERRICK, J. & AL., E. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. 269**,** 496-512.

FORTIER, L.-C. & SEKULOVIC, O. 2013. Importance of prophages to evolution and virulence of bacterial pathogens. *Virulence,* 4**,** 354-365.

FOURNIER, J. M. & QUILICI, M. L. 2007. [Cholera]. *Presse Med,* 36**,** 727-39.

FURRER, J. L., SANDERS, D. N., HOOK-BARNARD, I. G. & MCINTOSH, M. A. 2002. Export of the siderophore enterobactin in Escherichia coli: involvement of a 43 kDa membrane exporter. *Mol Microbiol,* 44**,** 1225-34.

GARRISON, E. & MARTH, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907*.

GARRISON, E., MARTH, G. 2012. Haplotype-based variant detection from short-read sequencing. *arXiv preprint,* 1207**,** 3907.

GAUDREAU, C., RATNAYAKE, R., PILON, P. A., GAGNON, S., ROGER, M. & LÉVESQUE, S. 2011. Ciprofloxacin-resistant Shigella sonnei among men who have sex with men, Canada, 2010. *Emerging infectious diseases,* 17**,** 1747-1750.

GAUTHIER, J., VINCENT, A. T., CHARETTE, S. J. & DEROME, N. 2018. A brief history of bioinformatics. *Briefings in Bioinformatics,* 20**,** 1981-1996.

GHATAK, S., KING, Z. A., SASTRY, A. & PALSSON, B. O. 2019. The y-ome defines the 35% of Escherichia coli genes that lack experimental evidence of function. *Nucleic Acids Research,* 47**,** 2446-2454.

GILBART, V. L., SIMMS, I., JENKINS, C., FUREGATO, M., GOBIN, M., OLIVER, I., HART, G., GILL, O. N. & HUGHES, G. 2015. Sex, drugs and smart phone applications: findings from semistructured interviews with men who have sex with men diagnosed with Shigella flexneri 3a in England and Wales. *Sex Transm Infect,* 91**,** 598-602.

GORGE, O., BENNETT, E. A., MASSILANI, D., DALIGAULT, J., PRUVOST, M., GEIGL, E. M. & GRANGE, T. 2016. Analysis of Ancient DNA in Microbial Ecology. *Methods Mol Biol,* 1399**,** 289-315.

GRANT, K., JENKINS, C., ARNOLD, C., GREEN, J. & ZAMBON, M. 2018. Implementing Pathogen Genomics - A Case Study. Public Health England.

GU, B., FAN, W., QIN, T., KONG, X., DONG, C., TAN, Z., CHEN, Y., XU, N., MA, P., BAO, C.-J. & QIAN, H. 2019. Existence of virulence genes in clinical Shigella sonnei isolates from Jiangsu Province of China: a multicenter study. *Annals of translational medicine,* 7**,** 305-305.

GÜNTHER, P., QUENTIN, D., AHMAD, S., SACHAR, K., GATSOGIANNIS, C., WHITNEY, J. C. & RAUNSER, S. 2022. Structure of a bacterial Rhs effector exported by the type VI secretion system. *PLOS Pathogens,* 18**,** e1010182.

GUPTA, S. D., LEE, B. T., CAMAKARIS, J. & WU, H. C. 1995. Identification of cutC and cutF (nlpE) genes involved in copper tolerance in Escherichia coli. *Journal of Bacteriology,* 177**,** 4207-4215.

HAFT, D. H., DICUCCIO, M., BADRETDIN, A., BROVER, V., CHETVERNIN, V., O'NEILL, K., LI, W., CHITSAZ, F., DERBYSHIRE, M. K., GONZALES, N. R., GWADZ, M., LU, F., MARCHLER, G. H., SONG, J. S., THANKI, N., YAMASHITA, R. A., ZHENG, C., THIBAUD-NISSEN, F., GEER, L. Y., MARCHLER-BAUER, A. & PRUITT, K. D. 2018. RefSeq: an update on prokaryotic genome annotation and curation. *Nucleic Acids Res,* 46**,** D851-d860.

HAHN-LOBMANN, S., STEPHAN, A., SCHULZ, S., SCHNEIDER, T., SHAVERSKYI, A., TUSE, D., GIRITCH, A. & GLEBA, Y. 2019. Colicins and Salmocins - New Classes of Plant-Made Non-antibiotic Food Antibacterials. *Front Plant Sci,* 10**,** 437.

HAHN-LÖBMANN, S., STEPHAN, A., SCHULZ, S., SCHNEIDER, T., SHAVERSKYI, A., TUSÉ, D., GIRITCH, A. & GLEBA, Y. 2019. Colicins and Salmocins - New Classes of Plant-Made Non-antibiotic Food Antibacterials. *Front Plant Sci,* 10**,** 437.

HALE, T. L. & KEUSCH, G. T. 1996. Shigella. *Medical Microbiology. 4th edition*.

HALL, R. J., WHELAN, F. J., CUMMINS, E. A., CONNOR, C., MCNALLY, A. & MCINERNEY, J. O. 2021. Gene-gene relationships in an Escherichia coli accessory genome are linked to function and mobility. *Microbial Genomics,* 7.

HAMMARLÖF, D. L., KRÖGER, C., OWEN, S. V., CANALS, R., LACHARME-LORA, L., WENNER, N., SCHAGER, A. E., WELLS, T. J., HENDERSON, I. R., WIGLEY, P., HOKAMP, K., FEASEY, N. A., GORDON, M. A. & HINTON, J. C. D. 2018. Role of a single noncoding nucleotide in the evolution of an epidemic African clade of Salmonella. *Proceedings of the National Academy of Sciences,* 115**,** E2614-E2623.

HAN, M. J. & LEE, S. Y. 2006. The Escherichia coli proteome: past, present, and future prospects. *Microbiol Mol Biol Rev,* 70**,** 362-439.

HARBOLA, A., NEGI, D., MANCHANDA, M. & KESHARWANI, R. K. 2022. Chapter 27 - Bioinformatics and biological data mining. *In:* SINGH, D. B. & PATHAK, R. K. (eds.) *Bioinformatics.* Academic Press.

HARRINGTON, A. T., HEARN, P. D., PICKING, W. L., BARKER, J. R., WESSEL, A. & PICKING, W. D. 2003. Structural characterization of the N terminus of IpaC from Shigella flexneri. *Infect Immun,* 71**,** 1255-64.

HARRISON, F., PAUL, J., MASSEY, R. C. & BUCKLING, A. 2008. Interspecific competition and siderophore-mediated cooperation in Pseudomonas aeruginosa. *Isme j,* 2**,** 49-55.

HASSANINASAB, A., HASHIMOTO, Y., TOMITA-YOKOTANI, K. & KOBAYASHI, M. 2011. Discovery of the curcumin metabolic pathway involving a unique enzyme in an intestinal microorganism. *Proc Natl Acad Sci U S A,* 108**,** 6615-20.

HAWKEY, J., MONK, J. M., BILLMAN-JACOBE, H., PALSSON, B. & HOLT, K. E. 2020. Impact of insertion sequences on convergent evolution of Shigella species. *PLOS Genetics,* 16**,** e1008931.

HAWKEY, J., PARANAGAMA, K., BAKER, K. S., BENGTSSON, R. J., WEILL, F.-X., THOMSON, N. R., BAKER, S., CERDEIRA, L., IQBAL, Z., HUNT, M., INGLE, D. J., DALLMAN, T. J., JENKINS, C., WILLIAMSON, D. A. & HOLT, K. E. 2021. Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, Shigella sonnei. *Nature Communications,* 12**,** 2684.

HIDER, R. C. & KONG, X. 2010. Chemistry and biology of siderophores. *Nat Prod Rep,* 27**,** 637-57.

HINIC, V., SETH-SMITH, H., STOCKLE, M., GOLDENBERGER, D. & EGLI, A. 2018. First report of sexually transmitted multi-drug resistant Shigella sonnei infections in Switzerland, investigated by whole genome sequencing. *Swiss Med Wkly,* 148**,** w14645.

HIRAMATSU, K. 2001. Vancomycin-resistant Staphylococcus aureus: a new model of antibiotic resistance. *The Lancet Infectious Diseases,* 1**,** 147-155.

HOLM, L. & LAAKSO, L. M. 2016. Dali server update. *Nucleic acids research,* 44**,** W351-W355.

HOLT, K. E., BAKER, S., WEILL, F.-X., HOLMES, E. C., KITCHEN, A., YU, J., SANGAL, V., BROWN, D. J., COIA, J. E., KIM, D. W., CHOI, S. Y., KIM, S. H., DA SILVEIRA, W. D., PICKARD, D. J., FARRAR, J. J., PARKHILL, J., DOUGAN, G. & THOMSON, N. R. 2012a. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics,* 44**,** 1056-1059.

HOLT, K. E., BAKER, S., WEILL, F. X., HOLMES, E. C., KITCHEN, A., YU, J., SANGAL, V., BROWN, D. J., COIA, J. E., KIM, D. W., CHOI, S. Y., KIM, S. H., DA SILVEIRA, W. D., PICKARD, D. J., FARRAR, J. J., PARKHILL, J., DOUGAN, G. & THOMSON, N. R. 2012b. Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat Genet,* 44**,** 1056-9.

HOLT, K. E., NGA, T. V. T., THANH, D. P., VINH, H., KIM, D. W., TRA, M. P. V., CAMPBELL, J. I., HOANG, N. V. M., VINH, N. T., MINH, P. V., THUY, C. T., NGA, T. T. T., THOMPSON, C., DUNG, T. T. N., NHU, N. T. K., VINH, P. V., TUYET, P. T. N., PHUC, H. L., LIEN, N. T. N., PHU, B. D., AI, N. T. T., TIEN, N. M., DONG, N., PARRY, C. M., HIEN, T. T., FARRAR, J. J., PARKHILL, J., DOUGAN, G., THOMSON, N. R. & BAKER, S. 2013. Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proceedings of the National Academy of Sciences,* 110**,** 17522-17527.

HOLT, K. E., WERTHEIM, H., ZADOKS, R. N., BAKER, S., WHITEHOUSE, C. A., DANCE, D., JENNEY, A., CONNOR, T. R., HSU, L. Y., SEVERIN, J., BRISSE, S., CAO, H., WILKSCH, J., GORRIE, C., SCHULTZ, M. B., EDWARDS, D. J., NGUYEN, K. V., NGUYEN, T. V., DAO, T. T., MENSINK, M., MINH, V. L., NHU, N. T. K., SCHULTSZ, C., KUNTAMAN, K., NEWTON, P. N., MOORE, C. E., STRUGNELL, R. A. & THOMSON, N. R. 2015. Genomic analysis of diversity, population structure,

virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proceedings of the National Academy of Sciences,* 112**,** E3574-E3581.

HORIYAMA, T. & NISHINO, K. 2014a. AcrB, AcrD, and MdtABC multidrug efflux systems are involved in enterobactin export in Escherichia coli. *PLoS One,* 9**,** e108642.

HORIYAMA, T. & NISHINO, K. 2014b. AcrB, AcrD, and MdtABC multidrug efflux systems are involved in enterobactin export in Escherichia coli. *PloS one,* 9**,** e108642-e108642.

HOVE-JENSEN, B., MCSORLEY, F. R. & ZECHEL, D. L. 2011. Physiological role of phnP-specified phosphoribosyl cyclic phosphodiesterase in catabolism of organophosphonic acids by the carbon-phosphorus lyase pathway. *J Am Chem Soc,* 133**,** 3617-24.

HOWELL, A. B., BOTTO, H., COMBESCURE, C., BLANC-POTARD, A.-B., GAUSA, L., MATSUMOTO, T., TENKE, P., SOTTO, A. & LAVIGNE, J.-P. 2010. Dosage effect on uropathogenic Escherichia coli anti-adhesion activity in urine following consumption of cranberry powder standardized for proanthocyanidin content: a multicentric randomized double blind study. *BMC infectious diseases,* 10**,** 1-11.

HU, D., LIU, B., FENG, L., DING, P., GUO, X., WANG, M., CAO, B., REEVES, P. R. & WANG, L. 2016. Origins of the current seventh cholera pandemic. *Proceedings of the National Academy of Sciences,* 113**,** E7730-E7739.

HUGHES, V. M. & DATTA, N. 1983. Conjugative plasmids in bacteria of the 'pre-antibiotic'era. *Nature,* 302**,** 725-726.

HUNT, M., BRADLEY, P., LAPIERRE, S. G., HEYS, S., THOMSIT, M., HALL, M. B., MALONE, K. M., WINTRINGER, P., WALKER, T. M., CIRILLO, D. M., COMAS, I., FARHAT, M. R., FOWLER, P., GARDY, J., ISMAIL, N., KOHL, T. A., MATHYS, V., MERKER, M., NIEMANN, S., OMAR, S. V., SINTCHENKO, V., SMITH, G., VAN SOOLINGEN, D., SUPPLY, P., TAHSEEN, S., WILCOX, M., ARANDJELOVIC, I., PETO, T. E. A., CROOK, D. W. & IQBAL, Z. 2019a. Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe. *Wellcome open research,* 4**,** 191-191.

HUNT, M., BRADLEY, P., LAPIERRE, S. G., HEYS, S., THOMSIT, M., HALL, M. B., MALONE, K. M., WINTRINGER, P., WALKER, T. M., CIRILLO, D. M., COMAS, I., FARHAT, M. R., FOWLER, P., GARDY, J., ISMAIL, N., KOHL, T. A., MATHYS, V., MERKER, M., NIEMANN, S., OMAR, S. V., SINTCHENKO, V., SMITH, G., VAN SOOLINGEN, D., SUPPLY, P., TAHSEEN, S., WILCOX, M., ARANDJELOVIC, I., PETO, T. E. A., CROOK, D. W. & IQBAL, Z. 2019b. Antibiotic resistance prediction for Mycobacterium tuberculosis from genome sequence data with Mykrobe. *Wellcome Open Res,* 4**,** 191.

ILBOUDO, P. G., HUANG, X. X., NGWIRA, B., MWANYUNGWE, A., MOGASALE, V., MENGEL, M. A., CAVAILLER, P., GESSNER, B. D. & LE GARGASSON, J.-B. 2017. Cost-of-illness of cholera to households and health facilities in rural Malawi. *PloS one,* 12**,** e0185041-e0185041.

INGLE, D. J., EASTON, M., VALCANIS, M., SEEMANN, T., KWONG, J. C., STEPHENS, N., CARTER, G. P., GONÇALVES DA SILVA, A., ADAMOPOULOS, J. & BAINES, S. L. 2019. Co-circulation of multidrug-resistant Shigella among men who have sex with men in Australia. *Clinical Infectious Diseases,* 69**,** 1535-1544.

INOUYE, M., DASHNOW, H., RAVEN, L. A., SCHULTZ, M. B., POPE, B. J., TOMITA, T., ZOBEL, J. & HOLT, K. E. 2014. SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med,* 6**,** 90.

JANSON, G., ZHANG, C., PRADO, M. G. & PAIARDINI, A. 2017. PyMod 2.0: improvements in protein sequence-structure analysis and homology modeling within PyMOL. *Bioinformatics,* 33**,** 444-446.

JAROSZEWSKI, L., KOSKA, L., SEDOVA, M. & GODZIK, A. 2014. PubServer: literature searches by homology. *Nucleic Acids Research,* 42**,** W430-W435.

JB, C. 1965. PROFFESSOR EGD MURRAY. AN APPRECIATION. *Canadian Medical Association Journal,* 92**,** 95-97.

JENNISON, A. V. & VERMA, N. K. 2004. Shigella flexneri infection: pathogenesis and vaccine development. *FEMS Microbiology Reviews,* 28**,** 43-58.

JONES, C. & STANLEY, J. 1992. Salmonella plasmids of the pre-antibiotic era. *J Gen Microbiol,* 138**,** 189-97.

JUMPER, J., EVANS, R., PRITZEL, A., GREEN, T., FIGURNOV, M., RONNEBERGER, O., TUNYASUVUNAKOOL, K., BATES, R., ŽÍDEK, A., POTAPENKO, A., BRIDGLAND, A., MEYER, C., KOHL, S. A. A., BALLARD, A. J., COWIE, A., ROMERA-PAREDES, B., NIKOLOV, S., JAIN, R., ADLER, J., BACK, T., PETERSEN, S., REIMAN, D., CLANCY, E., ZIELINSKI, M., STEINEGGER, M., PACHOLSKA, M., BERGHAMMER, T., BODENSTEIN, S., SILVER, D., VINYALS, O., SENIOR, A. W., KAVUKCUOGLU, K., KOHLI, P. & HASSABIS, D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature,* 596**,** 583-589.

JURADO-MARTÍN, I., SAINZ-MEJÍAS, M. & MCCLEAN, S. 2021. Pseudomonas aeruginosa: An Audacious Pathogen with an Adaptable Arsenal of Virulence Factors. *International Journal of Molecular Sciences,* 22**,** 3128.

KALLURI, P., CUMMINGS, K. C., ABBOTT, S., MALCOLM, G. B., HUTCHESON, K., BEALL, A., JOYCE, K., POLYAK, C., WOODWARD, D., CALDEIRA, R., RODGERS, F., MINTZ, E. D. & STROCKBINE, N. 2004. Epidemiological features of a newly described serotype of Shigella boydii. *Epidemiology and infection,* 132**,** 579-583.

KANIA, D. A., HAZEN, T. H., HOSSAIN, A., NATARO, J. P. & RASKO, D. A. 2016. Genome diversity of Shigella boydii. *Pathogens and Disease,* 74.

KHALIL, I. A., TROEGER, C., BLACKER, B. F., RAO, P. C., BROWN, A., ATHERLY, D. E., BREWER, T. G., ENGMANN, C. M., HOUPT, E. R. & KANG, G. 2018. Morbidity and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the Global Burden of Disease Study 1990–2016. *The Lancet infectious diseases,* 18**,** 1229-1240.

KLEMM, P. & SCHEMBRI, M. A. 2000. Bacterial adhesins: function and structure. *Int J Med Microbiol,* 290**,** 27-35.

KLEMM, P., TONG, S., NIELSEN, H. & CONWAY, T. 1996. The gntP gene of Escherichia coli involved in gluconate uptake. *J Bacteriol,* 178**,** 61-7.
KNAPP, M. & HOFREITER, M. 2010. Next Generation Sequencing of Ancient DNA: Requirements, Strategies and Perspectives. *Genes,* 1**,** 227-243.
KOESTLER, B. J., FISHER, C. R. & PAYNE, S. M. 2018. Formate Promotes Shigella Intercellular Spread and Virulence Gene Expression. *mBio,* 9.
KOHLER, P., SEIFFERT, S. N., KESSLER, S., RETTENMUND, G., LEMMENMEIER, E., QALLA WIDMER, L., NOLTE, O., SETH-SMITH, H. M. B., ALBRICH, W. C., BABOUEE FLURY, B., GARDIOL, C., HARBARTH, S., MÜNZER, T., SCHLEGEL, M., PETIGNAT, C., EGLI, A. & HÉQUET, D. 2022. Molecular Epidemiology and Risk Factors for Extended-Spectrum β-Lactamase–Producing Enterobacterales in Long-Term Care Residents. *Journal of the American Medical Directors Association,* 23**,** 475-481.e5.
KORRES, H., MAVRIS, M., MORONA, R., MANNING, P. A. & VERMA, N. K. 2005. Topological analysis of GtrA and GtrB proteins encoded by the serotype-converting cassette of Shigella flexneri. *Biochem Biophys Res Commun,* 328**,** 1252-60.
KOSKINIEMI, S., LAMOUREUX, J. G., NIKOLAKAKIS, K. C., T'KINT DE ROODENBEKE, C., KAPLAN, M. D., LOW, D. A. & HAYES, C. S. 2013. Rhs proteins from diverse bacteria mediate intercellular competition. *Proceedings of the National Academy of Sciences,* 110**,** 7032-7037.
KOTLOFF, K. L., NATARO, J. P., BLACKWELDER, W. C., NASRIN, D., FARAG, T. H., PANCHALINGAM, S., WU, Y., SOW, S. O., SUR, D., BREIMAN, R. F., FARUQUE, A. S. G., ZAIDI, A. K. M., SAHA, D., ALONSO, P. L., TAMBOURA, B., SANOGO, D., ONWUCHEKWA, U., MANNA, B., RAMAMURTHY, T., KANUNGO, S., OCHIENG, J. B., OMORE, R., OUNDO, J. O., HOSSAIN, A., DAS, S. K., AHMED, S., QURESHI, S., QUADRI, F., ADEGBOLA, R. A., ANTONIO, M., HOSSAIN, M. J., AKINSOLA, A., MANDOMANDO, I., NHAMPOSSA, T., ACÁCIO, S., BISWAS, K., O'REILLY, C. E., MINTZ, E. D., BERKELEY, L. Y., MUHSEN, K., SOMMERFELT, H., ROBINS-BROWNE, R. M. & LEVINE, M. M. 2013. Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study. *The Lancet,* 382**,** 209-222.
KOTLOFF, K. L., RIDDLE, M. S., PLATTS-MILLS, J. A., PAVLINAC, P. & ZAIDI, A. K. M. 2018. Shigellosis. *The Lancet,* 391**,** 801-812.
KOZLOV, A. M., DARRIBA, D., FLOURI, T., MOREL, B. & STAMATAKIS, A. 2019. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics,* 35**,** 4453-4455.
KRISTOFICOVA, I., VILHENA, C., BEHR, S. & JUNG, K. 2018. BtsT, a Novel and Specific Pyruvate/H(+) Symporter in Escherichia coli. *J Bacteriol,* 200.
KUBLER-KIELB, J., SCHNEERSON, R., MOCCA, C. & VINOGRADOV, E. 2008. The elucidation of the structure of the core part of the LPS from Plesiomonas

shigelloides serotype O17 expressing O-polysaccharide chain identical to the Shigella sonnei O-chain. *Carbohydrate Research,* 343**,** 3123-3127.
LAMPEL, K. A., FORMAL, S. B. & MAURELLI, A. T. 2018. A Brief History of Shigella. *EcoSal Plus,* 8**,** 10.1128/ecosalplus.ESP-0006-2017.
LAMPS, L. W. 2009. Shigella Species. *Surgical Pathology of the Gastrointestinal System: Bacterial, Fungal, Viral, and Parasitic Infections.* Springer.
LAZDUNSKI, C. J., BOUVERET, E., RIGAL, A., JOURNET, L., LLOUBES, R. & BENEDETTI, H. 1998. Colicin import into Escherichia coli cells. *J Bacteriol,* 180**,** 4993-5002.
LEES, J. A. & BENTLEY, S. D. 2016. Bacterial GWAS: not just gilding the lily. *Nature Reviews Microbiology,* 14**,** 406-406.
LEES, J. A., GALARDINI, M., BENTLEY, S. D., WEISER, J. N. & CORANDER, J. 2018. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics,* 34**,** 4310-4312.
LEES, J. A., VEHKALA, M., VÄLIMÄKI, N., HARRIS, S. R., CHEWAPREECHA, C., CROUCHER, N. J., MARTTINEN, P., DAVIES, M. R., STEER, A. C., TONG, S. Y. C., HONKELA, A., PARKHILL, J., BENTLEY, S. D. & CORANDER, J. 2016. Sequence element enrichment analysis to determine the genetic basis of bacterial phenotypes. *Nature Communications,* 7**,** 12797.
LETUNIC, I. & BORK, P. 2019. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research,* 47**,** W256-W259.
LETUNIC, I. & BORK, P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res,* 49**,** W293-W296.
LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.
LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009a. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.
LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009b. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.
LI, W. & GODZIK, A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics,* 22**,** 1658-9.
LI, W., O'NEILL, K. R., HAFT, D. H., DICUCCIO, M., CHETVERNIN, V., BADRETDIN, A., COULOURIS, G., CHITSAZ, F., DERBYSHIRE, M. K., DURKIN, A. S., GONZALES, N. R., GWADZ, M., LANCZYCKI, C. J., SONG, J. S., THANKI, N., WANG, J., YAMASHITA, R. A., YANG, M., ZHENG, C., MARCHLER-BAUER, A. & THIBAUD-NISSEN, F. 2021. RefSeq: expanding the Prokaryotic Genome Annotation Pipeline reach with protein family model curation. *Nucleic Acids Res,* 49**,** D1020-d1028.
LIN, H.-H., FILLOUX, A. & LAI, E.-M. 2020. Role of Recipient Susceptibility Factors During Contact-Dependent Interbacterial Competition. *Frontiers in Microbiology,* 11.

LIN, J., PENG, T., JIANG, L., NI, J.-Z., LIU, Q., CHEN, L. & ZHANG, Y. 2015. Comparative Genomics Reveals New Candidate Genes Involved in Selenium Metabolism in Prokaryotes. *Genome Biology and Evolution,* 7**,** 664-676.

LINDSAY, B., OUNDO, J., HOSSAIN, M. A., ANTONIO, M., TAMBOURA, B., WALKER, A. W., PAULSON, J. N., PARKHILL, J., OMORE, R., FARUQUE, A. S., DAS, S. K., IKUMAPAYI, U. N., ADEYEMI, M., SANOGO, D., SAHA, D., SOW, S., FARAG, T. H., NASRIN, D., LI, S., PANCHALINGAM, S., LEVINE, M. M., KOTLOFF, K., MAGDER, L. S., HUNGERFORD, L., SOMMERFELT, H., POP, M., NATARO, J. P. & STINE, O. C. 2015. Microbiota that affect risk for shigellosis in children in low-income countries. *Emerg Infect Dis,* 21**,** 242-50.

LIU, B., KNIREL, Y. A., FENG, L., PEREPELOV, A. V., SENCHENKOVA, S. N., WANG, Q., REEVES, P. R. & WANG, L. 2008a. Structure and genetics of Shigella O antigens. *FEMS Microbiol Rev,* 32**,** 627-53.

LIU, B., KNIREL, Y. A., FENG, L., PEREPELOV, A. V., SENCHENKOVA, S. Y. N., WANG, Q., REEVES, P. R. & WANG, L. 2008b. Structure and genetics of Shigella O antigens. *FEMS Microbiology Reviews,* 32**,** 627-653.

LIVIO, S., STROCKBINE, N. A., PANCHALINGAM, S., TENNANT, S. M., BARRY, E. M., MAROHN, M. E., ANTONIO, M., HOSSAIN, A., MANDOMANDO, I., OCHIENG, J. B., OUNDO, J. O., QURESHI, S., RAMAMURTHY, T., TAMBOURA, B., ADEGBOLA, R. A., HOSSAIN, M. J., SAHA, D., SEN, S., FARUQUE, A. S., ALONSO, P. L., BREIMAN, R. F., ZAIDI, A. K., SUR, D., SOW, S. O., BERKELEY, L. Y., O'REILLY, C. E., MINTZ, E. D., BISWAS, K., COHEN, D., FARAG, T. H., NASRIN, D., WU, Y., BLACKWELDER, W. C., KOTLOFF, K. L., NATARO, J. P. & LEVINE, M. M. 2014. Shigella isolates from the global enteric multicenter study inform vaccine development. *Clin Infect Dis,* 59**,** 933-41.

LOCHT, C., BERTIN, P., MENOZZI, F. D. & RENAULD, G. 1993. The filamentous haemagglutinin, a multifaceted adhesion produced by virulent Bordetella spp. *Mol Microbiol,* 9**,** 653-60.

LOCKE, R. K., GREIG, D. R., JENKINS, C., DALLMAN, T. J. & COWLEY, L. A. 2021. Acquisition and loss of CTX-M plasmids in Shigella species associated with MSM transmission in the UK. *Microb Genom,* 7.

LOGSDON, G. A., VOLLGER, M. R. & EICHLER, E. E. 2020. Long-read human genome sequencing and its applications. *Nature Reviews Genetics,* 21**,** 597-614.

LOOS, R. J. F. 2020. 15 years of genome-wide association studies and no signs of slowing down. *Nature Communications,* 11**,** 5900.

LUCCHINI, S., LIU, H., JIN, Q., HINTON, J. C. & YU, J. 2005. Transcriptional adaptation of Shigella flexneri during infection of macrophages and epithelial cells: insights into the strategies of a cytosolic bacterial pathogen. *Infect Immun,* 73**,** 88-102.

LUCK, S. N., TURNER, S. A., RAJAKUMAR, K., SAKELLARIS, H. & ADLER, B. 2001. Ferric dicitrate transport system (Fec) of Shigella flexneri 2a YSH6000 is encoded on a novel pathogenicity island carrying multiple antibiotic resistance genes. *Infect Immun,* 69**,** 6012-21.

LUNDRIGAN, M. D., KÖSTER, W. & KADNER, R. J. 1991. Transcribed sequences of the Escherichia coli btuB gene control its expression and regulation by vitamin B12. *Proceedings of the National Academy of Sciences,* 88**,** 1479-1483.

LYNCH, S. V., DIXON, L., BENOIT, M. R., BRODIE, E. L., KEYHAN, M., HU, P., ACKERLEY, D. F., ANDERSEN, G. L. & MATIN, A. 2007. Role of the rapA Gene in Controlling Antibiotic Resistance of Escherichia coli Biofilms. *Antimicrobial Agents and Chemotherapy,* 51**,** 3650-3658.

MACINTYRE, D. L., MIYATA, S. T., KITAOKA, M. & PUKATZKI, S. 2010. The Vibrio cholerae type VI secretion system displays antimicrobial properties. *Proc Natl Acad Sci U S A,* 107**,** 19520-4.

MALAKA DE SILVA, P., STENHOUSE, G. E., BLACKWELL, G. A., BENGTSSON, R. J., JENKINS, C., HALL, J. P. J. & BAKER, K. S. 2022. A tale of two plasmids: contributions of plasmid associated phenotypes to epidemiological success among Shigella. *Proc Biol Sci,* 289**,** 20220581.

MALEKIAN, N., AGRAWAL, A. A., BERENDONK, T. U., AL-FATLAWI, A. & SCHROEDER, M. 2022. A genome-wide scan of wastewater E. coli for genes under positive selection: focusing on mechanisms of antibiotic resistance. *Scientific Reports,* 12**,** 8037.

MANSON, M. D. & YANOFSKY, C. 1976. Naturally occurring sites within the Shigella dysenteriae tryptophan operon severely limit tryptophan biosynthesis. *J Bacteriol,* 126**,** 668-78.

MANSON-BAHR, P. H. 1942. Dysentery and Diarrhoea in Wartime. *British medical journal,* 2**,** 346-348.

MANTERE, T., KERSTEN, S. & HOISCHEN, A. 2019. Long-Read Sequencing Emerging in Medical Genetics. *Frontiers in Genetics,* 10.

MARGULIES, M., EGHOLM, M., ALTMAN, W. E., ATTIYA, S., BADER, J. S., BEMBEN, L. A., BERKA, J., BRAVERMAN, M. S., CHEN, Y.-J., CHEN, Z., DEWELL, S. B., DU, L., FIERRO, J. M., GOMES, X. V., GODWIN, B. C., HE, W., HELGESEN, S., HO, C. H., IRZYK, G. P., JANDO, S. C., ALENQUER, M. L. I., JARVIE, T. P., JIRAGE, K. B., KIM, J.-B., KNIGHT, J. R., LANZA, J. R., LEAMON, J. H., LEFKOWITZ, S. M., LEI, M., LI, J., LOHMAN, K. L., LU, H., MAKHIJANI, V. B., MCDADE, K. E., MCKENNA, M. P., MYERS, E. W., NICKERSON, E., NOBILE, J. R., PLANT, R., PUC, B. P., RONAN, M. T., ROTH, G. T., SARKIS, G. J., SIMONS, J. F., SIMPSON, J. W., SRINIVASAN, M., TARTARO, K. R., TOMASZ, A., VOGT, K. A., VOLKMER, G. A., WANG, S. H., WANG, Y., WEINER, M. P., YU, P., BEGLEY, R. F. & ROTHBERG, J. M. 2005. Genome sequencing in microfabricated high-density picolitre reactors. *Nature,* 437**,** 376-380.

MARTINEZ-ARGUDO, I. & BLOCKER, A. J. 2010. The Shigella T3SS needle transmits a signal for MxiC release, which controls secretion of effectors. *Mol Microbiol,* 78**,** 1365-78.

MATHERS, A. J., PEIRANO, G. & PITOUT, J. D. 2015. The role of epidemic resistance plasmids and international high-risk clones in the spread of multidrug-resistant Enterobacteriaceae. *Clinical microbiology reviews,* 28**,** 565-591.

MATTOCK, E. & BLOCKER, A. J. 2017. How Do the Virulence Factors of Shigella Work Together to Cause Disease? *Frontiers in Cellular and Infection Microbiology,* 7.
MCIVER, C. J., WHITE, P. A., JONES, L. A., KARAGIANNIS, T., HARKNESS, J., MARRIOTT, D. & RAWLINSON, W. D. 2002. Epidemic Strains of Shigella sonnei Biotype g Carrying Integrons. *Journal of Clinical Microbiology,* 40**,** 1538-1540.
MCVICKER, G. & TANG, C. M. 2016. Deletion of toxin–antitoxin systems in the evolution of Shigella sonnei as a host-adapted pathogen. *Nature Microbiology,* 2**,** 1-8.
MEDINA, V., PONTAROLLO, R., GLAESKE, D., TABEL, H. & GOLDIE, H. 1990. Sequence of the pckA gene of Escherichia coli K-12: relevance to genetic and allosteric regulation and homology of E. coli phosphoenolpyruvate carboxykinase with the enzymes from Trypanosoma brucei and Saccharomyces cerevisiae. *J Bacteriol,* 172**,** 7151-6.
MEHMOOD, M. A., SEHAR, U. & AHMAD, N. 2014. Use of bioinformatics tools in different spheres of life sciences. *Journal of Data Mining in Genomics & Proteomics,* 5**,** 1.
MELTON-CELSA, A. R. 2014. Shiga Toxin (Stx) Classification, Structure, and Function. *Microbiology spectrum,* 2**,** 10.1128/microbiolspec.EHEC-0024-2013-2013.
MENON, N. K., ROBBINS, J., PECK, H. D., JR., CHATELUS, C. Y., CHOI, E. S. & PRZYBYLA, A. E. 1990. Cloning and sequencing of a putative Escherichia coli [NiFe] hydrogenase-1 operon containing six open reading frames. *J Bacteriol,* 172**,** 1969-77.
MEY, A. R., GÓMEZ-GARZÓN, C. & PAYNE, S. M. 2021. Iron Transport and Metabolism in Escherichia, Shigella, and Salmonella. *EcoSal Plus,* 9**,** eESP-0034-2020.
MILADI, H., MILI, D., SLAMA, R. B., ZOUARI, S., AMMAR, E. & BAKHROUF, A. 2016. Antibiofilm formation and anti-adhesive property of three mediterranean essential oils against a foodborne pathogen Salmonella strain. *Microbial pathogenesis,* 93**,** 22-31.
MILLER, C. & COHEN, S. N. 1999. Separate roles of Escherichia coli replication proteins in synthesis and partitioning of pSC101 plasmid DNA. *J Bacteriol,* 181**,** 7552-7.
MITRA, R., MCKENZIE, G. J., YI, L., LEE, C. A. & CRAIG, N. L. 2010. Characterization of the TnsD-attTn7 complex that promotes site-specific insertion of Tn7. *Mobile DNA,* 1**,** 18.
MOGASALE, V., NGOGOYO, S. M. & MOGASALE, V. V. 2021. Model-based estimation of the economic burden of cholera in Africa. *BMJ Open,* 11**,** e044615.
MØLLER, T. S. B., OVERGAARD, M., NIELSEN, S. S., BORTOLAIA, V., SOMMER, M. O. A., GUARDABASSI, L. & OLSEN, J. E. 2016. Relation between tetR and tetA expression in tetracycline resistant Escherichia coli. *BMC Microbiology,* 16**,** 39.
MONJARÁS FERIA, J. & VALVANO, M. A. 2020. An Overview of Anti-Eukaryotic T6SS Effectors. *Frontiers in Cellular and Infection Microbiology,* 10.
MONTMINY, S. W., KHAN, N., MCGRATH, S., WALKOWICZ, M. J., SHARP, F., CONLON, J. E., FUKASE, K., KUSUMOTO, S., SWEET, C. & MIYAKE, K. 2006. Virulence factors

of Yersinia pestis are overcome by a strong lipopolysaccharide response. *Nature immunology,* 7**,** 1066-1073.

MOODY, G. 2004. *Digital code of life: how bioinformatics is revolutionizing science, medicine, and business*, John Wiley & Sons.

MORENO-MINGORANCE, A., ESPINAL, P., RODRIGUEZ, V., GOTERRIS, L., FABREGA, A., SERRA-PLADEVALL, J., BARBERA, M. J., ALBERNY, M., MARTIN-GONZALEZ, H., CORNEJO-SANCHEZ, T., ARMAS, M., MIR-CROS, A., RAVENTOS, A., VINADO, B., PUMAROLA, T., LARROSA, M. N. & GONZALEZ-LOPEZ, J. J. 2021. Circulation of multi-drug-resistant Shigella sonnei and Shigella flexneri among men who have sex with men in Barcelona, Spain, 2015-2019. *Int J Antimicrob Agents,* 58**,** 106378.

MURRAY, C. J. L., IKUTA, K. S., SHARARA, F., SWETSCHINSKI, L., ROBLES AGUILAR, G., GRAY, A., HAN, C., BISIGNANO, C., RAO, P., WOOL, E., JOHNSON, S. C., BROWNE, A. J., CHIPETA, M. G., FELL, F., HACKETT, S., HAINES-WOODHOUSE, G., KASHEF HAMADANI, B. H., KUMARAN, E. A. P., MCMANIGAL, B., AGARWAL, R., AKECH, S., ALBERTSON, S., AMUASI, J., ANDREWS, J., ARAVKIN, A., ASHLEY, E., BAILEY, F., BAKER, S., BASNYAT, B., BEKKER, A., BENDER, R., BETHOU, A., BIELICKI, J., BOONKASIDECHA, S., BUKOSIA, J., CARVALHEIRO, C., CASTAÑEDA-ORJUELA, C., CHANSAMOUTH, V., CHAURASIA, S., CHIURCHIÙ, S., CHOWDHURY, F., COOK, A. J., COOPER, B., CRESSEY, T. R., CRIOLLO-MORA, E., CUNNINGHAM, M., DARBOE, S., DAY, N. P. J., DE LUCA, M., DOKOVA, K., DRAMOWSKI, A., DUNACHIE, S. J., ECKMANNS, T., EIBACH, D., EMAMI, A., FEASEY, N., FISHER-PEARSON, N., FORREST, K., GARRETT, D., GASTMEIER, P., GIREF, A. Z., GREER, R. C., GUPTA, V., HALLER, S., HASELBECK, A., HAY, S. I., HOLM, M., HOPKINS, S., IREGBU, K. C., JACOBS, J., JAROVSKY, D., JAVANMARDI, F., KHORANA, M., KISSOON, N., KOBEISSI, E., KOSTYANEV, T., KRAPP, F., KRUMKAMP, R., KUMAR, A., KYU, H. H., LIM, C., LIMMATHUROTSAKUL, D., LOFTUS, M. J., LUNN, M., MA, J., MTURI, N., MUNERA-HUERTAS, T., MUSICHA, P., MUSSI-PINHATA, M. M., NAKAMURA, T., NANAVATI, R., NANGIA, S., NEWTON, P., NGOUN, C., NOVOTNEY, A., NWAKANMA, D., OBIERO, C. W., OLIVAS-MARTINEZ, A., OLLIARO, P., OOKO, E., et al. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet,* 399**,** 629-655.

MUTHUIRULANDI SETHUVEL, D. P., DEVANGA RAGUPATHI, N. K., ANANDAN, S. & VEERARAGHAVAN, B. 2017. Update on: Shigella new serogroups/serotypes and their antimicrobial resistance. *Letters in Applied Microbiology,* 64**,** 8-18.

MUTHURAMALINGAM, M., WHITTIER, S. K., PICKING, W. L. & PICKING, W. D. 2021. The Shigella Type III Secretion System: An Overview from Top to Bottom. *Microorganisms,* 9**,** 451.

NEWTON, H. J., PEARSON, J. S., BADEA, L., KELLY, M., LUCAS, M., HOLLOWAY, G., WAGSTAFF, K. M., DUNSTONE, M. A., SLOAN, J. & WHISSTOCK, J. C. 2010. The type III effectors NleE and NleB from enteropathogenic E. coli and OspZ from

Shigella block nuclear translocation of NF-κB p65. *PLoS pathogens,* 6**,** e1000898.

NIEBUHR, K., JOUIHRI, N., ALLAOUI, A., GOUNON, P., SANSONETTI, P. J. & PARSOT, C. 2000. IpgD, a protein secreted by the type III secretion machinery of Shigella flexneri, is chaperoned by IpgE and implicated in entry focus formation. *Mol Microbiol,* 38**,** 8-19.

NIJKAMP, H. J., DE LANG, R., STUITJE, A. R., VAN DEN ELZEN, P. J., VELTKAMP, E. & VAN PUTTEN, A. J. 1986. The complete nucleotide sequence of the bacteriocinogenic plasmid CloDF13. *Plasmid,* 16**,** 135-60.

NISA, I., QASIM, M., YASIN, N., ULLAH, R. & ALI, A. 2020. Shigella flexneri: an emerging pathogen. *Folia Microbiologica,* 65**,** 275-291.

NIYOGI, S. K., VARGAS, M. & VILA, J. 2004. Prevalence of the sat, set and sen genes among diverse serotypes of Shigella flexneri strains isolated from patients with acute diarrhoea. *Clin Microbiol Infect,* 10**,** 574-6.

NJAMKEPO, E., FAWAL, N., TRAN-DIEN, A., HAWKEY, J., STROCKBINE, N., JENKINS, C., TALUKDER, K. A., BERCION, R., KULESHOV, K., KOLÍNSKÁ, R., RUSSELL, J. E., KAFTYREVA, L., ACCOU-DEMARTIN, M., KARAS, A., VANDENBERG, O., MATHER, A. E., MASON, C. J., PAGE, A. J., RAMAMURTHY, T., BIZET, C., GAMIAN, A., CARLE, I., SOW, A. G., BOUCHIER, C., WESTER, A. L., LEJAY-COLLIN, M., FONKOUA, M.-C., LE HELLO, S., BLASER, M. J., JERNBERG, C., RUCKLY, C., MÉRENS, A., PAGE, A.-L., ASLETT, M., ROGGENTIN, P., FRUTH, A., DENAMUR, E., VENKATESAN, M., BERCOVIER, H., BODHIDATTA, L., CHIOU, C.-S., CLERMONT, D., COLONNA, B., EGOROVA, S., PAZHANI, G. P., EZERNITCHI, A. V., GUIGON, G., HARRIS, S. R., IZUMIYA, H., KORZENIOWSKA-KOWAL, A., LUTYŃSKA, A., GOUALI, M., GRIMONT, F., LANGENDORF, C., MAREJKOVÁ, M., PETERSON, L. A. M., PEREZ-PEREZ, G., NGANDJIO, A., PODKOLZIN, A., SOUCHE, E., MAKAROVA, M., SHIPULIN, G. A., YE, C., ŽEMLIČKOVÁ, H., HERPAY, M., GRIMONT, P. A. D., PARKHILL, J., SANSONETTI, P., HOLT, K. E., BRISSE, S., THOMSON, N. R. & WEILL, F.-X. 2016a. Global phylogeography and evolutionary history of Shigella dysenteriae type 1. *Nature Microbiology,* 1**,** 16027.

NJAMKEPO, E., FAWAL, N., TRAN-DIEN, A., HAWKEY, J., STROCKBINE, N., JENKINS, C., TALUKDER, K. A., BERCION, R., KULESHOV, K., KOLÍNSKÁ, R., RUSSELL, J. E., KAFTYREVA, L., ACCOU-DEMARTIN, M., KARAS, A., VANDENBERG, O., MATHER, A. E., MASON, C. J., PAGE, A. J., RAMAMURTHY, T., BIZET, C., GAMIAN, A., CARLE, I., SOW, A. G., BOUCHIER, C., WESTER, A. L., LEJAY-COLLIN, M., FONKOUA, M.-C., LE HELLO, S., BLASER, M. J., JERNBERG, C., RUCKLY, C., MÉRENS, A., PAGE, A.-L., ASLETT, M., ROGGENTIN, P., FRUTH, A., DENAMUR, E., VENKATESAN, M., BERCOVIER, H., BODHIDATTA, L., CHIOU, C.-S., CLERMONT, D., COLONNA, B., EGOROVA, S., PAZHANI, G. P., EZERNITCHI, A. V., GUIGON, G., HARRIS, S. R., IZUMIYA, H., KORZENIOWSKA-KOWAL, A., LUTYŃSKA, A., GOUALI, M., GRIMONT, F., LANGENDORF, C., MAREJKOVÁ, M., PETERSON, L. A. M., PEREZ-PEREZ, G., NGANDJIO, A., PODKOLZIN, A., SOUCHE, E., MAKAROVA, M., SHIPULIN, G. A., YE, C., ŽEMLIČKOVÁ, H., HERPAY, M., GRIMONT, P. A. D.,

PARKHILL, J., SANSONETTI, P., HOLT, K. E., BRISSE, S., THOMSON, N. R. & WEILL, F.-X. 2016b. Global phylogeography and evolutionary history of Shigella dysenteriae type 1. *Nature Microbiology,* 1**,** 16027.

NOBELMANN, B. & LENGELER, J. W. 1996. Molecular analysis of the gat genes from Escherichia coli and of their roles in galactitol transport and metabolism. *Journal of Bacteriology,* 178**,** 6790-6795.

NOGALES, J. & GARMENDIA, J. 2022. Bacterial metabolism and pathogenesis intimate intertwining: time for metabolic modelling to come into action. *Microbial Biotechnology,* 15**,** 95-102.

NOTTI, R. Q. & STEBBINS, C. E. 2016. The Structure and Function of Type III Secretion Systems. *Microbiol Spectr,* 4.

NYGREN, B. L., SCHILLING, K. A., BLANTON, E. M., SILK, B. J., COLE, D. J. & MINTZ, E. D. 2013. Foodborne outbreaks of shigellosis in the USA, 1998–2008. *Epidemiology and Infection,* 141**,** 233-241.

NYMAN, K., NAKAMURA, K., OHTSUBO, H. & OHTSUBO, E. 1981. Distribution of the insertion sequence IS1 in Gram-negative bacteria. *Nature,* 289**,** 609-612.

O'NEIL, J. 2014. Tackling a crisis for the health and wealth of nations. *World Health Organization*.

OECD 2018. *Stemming the Superbug Tide*.

OFEK, I., HASTY, D. L. & SHARON, N. 2003. Anti-adhesion therapy of bacterial diseases: prospects and problems. *FEMS Immunology & Medical Microbiology,* 38**,** 181-191.

OHTSUBO, H., NYMAN, K., DOROSZKIEWICZ, W. & OHTSUBO, E. 1981. Multiple copies of iso-insertion sequences of IS1 in Shigella dysenteriae chromosome. *Nature,* 292**,** 640-643.

PACIELLO, I., SILIPO, A., LEMBO-FAZIO, L., CURCURÙ, L., ZUMSTEG, A., NOËL, G., CIANCARELLA, V., STURIALE, L., MOLINARO, A. & BERNARDINI, M. L. 2013. Intracellular Shigella remodels its LPS to dampen the innate immune recognition and evade inflammasome activation. *Proc Natl Acad Sci U S A,* 110**,** E4345-54.

PAGE, A. J., CUMMINS, C. A., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T., FOOKES, M., FALUSH, D., KEANE, J. A. & PARKHILL, J. 2015a. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics,* 31**,** 3691-3.

PAGE, A. J., CUMMINS, C. A., KEANE, J. A., PARKHILL, J., FOOKES, M., HUNT, M., WONG, V. K., REUTER, S., HOLDEN, M. T. G. & FALUSH, D. 2015b. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics,* 31**,** 3691-3693.

PAGE, A. J., TAYLOR, B., DELANEY, A. J., SOARES, J., SEEMANN, T., KEANE, J. A. & HARRIS, S. R. 2016. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom,* 2**,** e000056.

PANDURANGAN, A. P., OCHOA-MONTAÑO, B., ASCHER, D. B. & BLUNDELL, T. L. 2017. SDM: a server for predicting effects of mutations on protein stability. *Nucleic Acids Research,* 45**,** W229-W235.

PASSALACQUA, K. D., CHARBONNEAU, M. E. & O'RIORDAN, M. X. D. 2016. Bacterial Metabolism Shapes the Host-Pathogen Interface. *Microbiol Spectr,* 4.

PAVONCELLO, V., BARRAS, F. & BOUVERET, E. 2022. Degradation of Exogenous Fatty Acids in Escherichia coli. *Biomolecules,* 12**,** 1019.

PAYNE, S. M., WYCKOFF, E. E., MURPHY, E. R., OGLESBY, A. G., BOULETTE, M. L. & DAVIES, N. M. L. 2006. Iron and Pathogenesis of Shigella: Iron Acquisition in the Intracellular Environment. *Biometals,* 19**,** 173-180.

PETERS, J. E. & CRAIG, N. L. 2001. Tn7 recognizes transposition target structures associated with DNA replication using the DNA-binding protein TnsE. *Genes Dev,* 15**,** 737-47.

PICKING, W. L., COYE, L., OSIECKI, J. C., SERFIS, A. B., SCHAPER, E. & PICKING, W. D. 2001. Identification of functional regions within invasion plasmid antigen C (IpaC) of Shigella flexneri. *Molecular Microbiology,* 39**,** 100-111.

PIDDOCK, L. J. 2006. Clinically relevant chromosomally encoded multidrug resistance efflux pumps in bacteria. *Clinical microbiology reviews,* 19**,** 382-402.

PILLA, G., MCVICKER, G. & TANG, C. M. 2017. Genetic plasticity of the Shigella virulence plasmid is mediated by intra- and inter-molecular events between insertion sequences. *PLOS Genetics,* 13**,** e1007014.

PLANET, P. J., LARUSSA, S. J., DANA, A., SMITH, H., XU, A., RYAN, C., UHLEMANN, A. C., BOUNDY, S., GOLDBERG, J., NARECHANIA, A., KULKARNI, R., RATNER, A. J., GEOGHEGAN, J. A., KOLOKOTRONIS, S. O. & PRINCE, A. 2013. Emergence of the epidemic methicillin-resistant Staphylococcus aureus strain USA300 coincides with horizontal transfer of the arginine catabolic mobile element and speG-mediated adaptations for survival on skin. *mBio,* 4**,** e00889-13.

PLETZER, D. & WEINGART, H. 2014. Characterization and regulation of the Resistance-Nodulation-Cell Division-type multidrug efflux pumps MdtABC and MdtUVW from the fire blight pathogen Erwinia amylovora. *BMC Microbiology,* 14**,** 185.

PODOLSKY, S. H. 2018. The evolving response to antibiotic resistance (1945–2018). *Palgrave Communications,* 4**,** 124.

POP, M., WALKER, A. W., PAULSON, J., LINDSAY, B., ANTONIO, M., HOSSAIN, M. A., OUNDO, J., TAMBOURA, B., MAI, V., ASTROVSKAYA, I., CORRADA BRAVO, H., RANCE, R., STARES, M., LEVINE, M. M., PANCHALINGAM, S., KOTLOFF, K., IKUMAPAYI, U. N., EBRUKE, C., ADEYEMI, M., AHMED, D., AHMED, F., ALAM, M. T., AMIN, R., SIDDIQUI, S., OCHIENG, J. B., OUMA, E., JUMA, J., MAILU, E., OMORE, R., MORRIS, J. G., BREIMAN, R. F., SAHA, D., PARKHILL, J., NATARO, J. P. & STINE, O. C. 2014. Diarrhea in young children from low-income countries leads to large-scale alterations in intestinal microbiota composition. *Genome Biol,* 15**,** R76.

PORE, D., MAHATA, N., PAL, A. & CHAKRABARTI, M. K. 2011. Outer membrane protein A (OmpA) of Shigella flexneri 2a, induces protective immune response in a mouse model. *PLoS One,* 6**,** e22663.

PORTER, C. K., THURA, N., RANALLO, R. T. & RIDDLE, M. S. 2013. The Shigella human challenge model. *Epidemiol Infect,* 141**,** 223-32.

PRICE-WHELAN, A., DIETRICH, L. E. P. & NEWMAN, D. K. 2006. Rethinking 'secondary' metabolism: physiological roles for phenazine antibiotics. *Nature Chemical Biology,* 2**,** 71-78.

PROBER, J. M., TRAINOR, G. L., DAM, R. J., HOBBS, F. W., ROBERTSON, C. W., ZAGURSKY, R. J., COCUZZA, A. J., JENSEN, M. A. & BAUMEISTER, K. 1987. A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science,* 238**,** 336-341.

PUGSLEY, A. P. 1988. The immunity and lysis genes of ColN plasmid pCHAP4. *Mol Gen Genet,* 211**,** 335-41.

PURCELL, S., NEALE, B., TODD-BROWN, K., THOMAS, L., FERREIRA, M. A. R., BENDER, D., MALLER, J., SKLAR, P., DE BAKKER, P. I. W., DALY, M. J. & SHAM, P. C. 2007. PLINK: a tool set for whole-genome association and population-based linkage analyses. *American journal of human genetics,* 81**,** 559-575.

RABAA, M. A., THANH, D. P., DE LAPPE, N., CORMICAN, M., VALCANIS, M., HOWDEN, B. P., WANGCHUK, S., BODHIDATTA, L., MASON, C. J. & THI, T. N. N. 2016. South Asia as a reservoir for the global spread of ciprofloxacin resistant Shigella sonnei. *bioRxiv***,** 041327.

RAM, P., CRUMP, J., GUPTA, S., MILLER, M. & MINTZ, E. 2008. Part II. Analysis of data gaps pertaining to Shigella infections in low and medium human development index countries, 1984–2005. *Epidemiology & Infection,* 136**,** 577-603.

REDDICK, L. E. & ALTO, N. M. 2014. Bacteria fighting back: how pathogens target and subvert the host innate immune system. *Molecular cell,* 54**,** 321-328.

REGION, A., REGION, S.-E. A., REGION, E. M. & REGION, W. P. Global action plan on antimicrobial resistance.

RELLER, M. E., NELSON, J. M., MØLBAK, K., ACKMAN, D. M., SCHOONMAKER-BOPP, D. J., ROOT, T. P. & MINTZ, E. D. 2006. A large, multiple-restaurant outbreak of infection with Shigella flexneri serotype 2a traced to tomatoes. *Clinical infectious diseases,* 42**,** 163-169.

REW, V., MOOK, P., TRIENEKENS, S., BAKER, K. S., DALLMAN, T. J., JENKINS, C., CROOK, P. D. & THOMSON, N. R. 2018. Whole-genome sequencing revealed concurrent outbreaks of shigellosis in the English Orthodox Jewish Community caused by multiple importations of Shigella sonnei from Israel. *Microb Genom,* 4.

RIEGMAN, N., KUSTERS, R., VAN VEGGEL, H., BERGMANS, H., VAN BERGEN EN HENEGOUWEN, P., HACKER, J. & VAN DIE, I. 1990. F1C fimbriae of a uropathogenic Escherichia coli strain: genetic and functional organization of the foc gene cluster and identification of minor subunits. *J Bacteriol,* 172**,** 1114-20.

RILEY, M. A. 1993. Molecular mechanisms of colicin evolution. *Mol Biol Evol,* 10**,** 1380-95.

RODRIGUES, C. H. M., PIRES, D. E. V. & ASCHER, D. B. 2021. DynaMut2: Assessing changes in stability and flexibility upon single and multiple point missense mutations. *Protein Sci,* 30**,** 60-69.

ROHMER, L., JACOBS, M. A., BRITTNACHER, M. J., FONG, C., HAYDEN, H. S., HOCQUET, D., WEISS, E. J., RADEY, M., GERMANI, Y., TALUKDER, K. A., HAGER, A. J., KEMNER, J. M., SIMS-DAY, E. H., MATAMOUROS, S., HAGER, K. R. & MILLER, S. I. 2014. Genomic analysis of the emergence of 20th century epidemic dysentery. *BMC Genomics,* 15**,** 355.

RUNYEN-JANECKY, L. J., REEVES, S. A., GONZALES, E. G. & PAYNE, S. M. 2003. Contribution of the Shigella flexneri Sit, Iuc, and Feo iron acquisition systems to iron acquisition in vitro and in cultured cells. *Infect Immun,* 71**,** 1919-28.

RUTHERFORD, S. T., LEMKE, J. J., VRENTAS, C. E., GAAL, T., ROSS, W. & GOURSE, R. L. 2007. Effects of DksA, GreA, and GreB on transcription initiation: insights into the mechanisms of factors that bind in the secondary channel of RNA polymerase. *J Mol Biol,* 366**,** 1243-57.

SABER, M. M. & SHAPIRO, B. J. 2020. Benchmarking bacterial genome-wide association study methods using simulated genomes and phenotypes. *Microb Genom,* 6.

SACK, D. A., HOQUE, A. T., HUQ, A. & ETHERIDGE, M. 1994. Is protection against shigellosis induced by natural infection with Plesiomonas shigelloides? *Lancet,* 343**,** 1413-5.

SAHINTOTH, M., FRILLINGOS, S., LENGELER, J. W. & KABACK, H. R. 1995. Active Transport by the CscB Permease in Escherichia coli K-12. *Biochemical and Biophysical Research Communications,* 208**,** 1116-1123.

SAHL, J. W., MORRIS, C. R., EMBERGER, J., FRASER, C. M., OCHIENG, J. B., JUMA, J., FIELDS, B., BREIMAN, R. F., GILMOUR, M., NATARO, J. P. & RASKO, D. A. 2015. Defining the phylogenomics of Shigella species: a pathway to diagnostics. *Journal of clinical microbiology,* 53**,** 951-960.

SAKELLARIS, H., HANNINK, N. K., RAJAKUMAR, K., BULACH, D., HUNT, M., SASAKAWA, C. & ADLER, B. 2000. Curli loci of Shigella spp. *Infect Immun,* 68**,** 3780-3.

SAN, J. E., BAICHOO, S., KANZI, A., MOOSA, Y., LESSELLS, R., FONSECA, V., MOGAKA, J., POWER, R. & DE OLIVEIRA, T. 2019. Current Affairs of Microbial Genome-Wide Association Studies: Approaches, Bottlenecks and Analytical Pitfalls. *Front Microbiol,* 10**,** 3119.

SANADA, T., KIM, M., MIMURO, H., SUZUKI, M., OGAWA, M., OYAMA, A., ASHIDA, H., KOBAYASHI, T., KOYAMA, T. & NAGAI, S. 2012. The Shigella flexneri effector OspI deamidates UBC13 to dampen the inflammatory response. *Nature,* 483**,** 623-626.

SANDERS, M. E., NORCROSS, E. W., ROBERTSON, Z. M., MOORE, Q. C., III, FRATKIN, J. & MARQUART, M. E. 2011. The Streptococcus pneumoniae Capsule Is Required for Full Virulence in Pneumococcal Endophthalmitis. *Investigative Ophthalmology & Visual Science,* 52**,** 865-872.

SANGER, F., NICKLEN, S. & COULSON, A. R. 1977. DNA sequencing with chain-terminating inhibitors. *Proc Natl Acad Sci U S A,* 74**,** 5463-7.

SANGER, F. & THOMPSON, E. 1953. The amino-acid sequence in the glycyl chain of insulin. 1. The identification of lower peptides from partial hydrolysates. *Biochemical Journal,* 53**,** 353.
SANSONETTI, P., KOPECKO, D. & FORMAL, S. 1982. Involvement of a plasmid in the invasive ability of Shigella flexneri. *Infection and immunity,* 35**,** 852-860.
SANSONETTI, P. J. & PHALIPON, A. M cells as ports of entry for enteroinvasive pathogens: mechanisms of interaction, consequences for the disease process. Seminars in immunology, 1999. Elsevier, 193-203.
SARGENT, F., STANLEY, N. R., BERKS, B. C. & PALMER, T. 1999. Sec-independent protein translocation in Escherichia coli: a distinct and pivotal role for the TatB protein. *Journal of Biological Chemistry,* 274**,** 36073-36082.
SCHMITT, M. P. & PAYNE, S. M. 1988. Genetics and regulation of enterobactin genes in Shigella flexneri. *J Bacteriol,* 170**,** 5579-87.
SCHROVEN, K., AERTSEN, A. & LAVIGNE, R. 2020. Bacteriophages as drivers of bacterial virulence and their potential for biotechnological exploitation. *FEMS Microbiology Reviews,* 45.
SCHWARZ, S., WEST, T. E., BOYER, F., CHIANG, W.-C., CARL, M. A., HOOD, R. D., ROHMER, L., TOLKER-NIELSEN, T., SKERRETT, S. J. & MOUGOUS, J. D. 2010. Burkholderia type VI secretion systems have distinct roles in eukaryotic and bacterial cell interactions. *PLoS pathogens,* 6**,** e1001068.
SEEMANN, T. Abricate. Github https://github.com/tseemann/abricate.
SEEMANN, T. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics,* 30**,** 2068-9.
SEN, T. & VERMA, N. K. 2020. Functional Annotation and Curation of Hypothetical Proteins Present in A Newly Emerged Serotype 1c of Shigella flexneri: Emphasis on Selecting Targets for Virulence and Vaccine Design Studies. *Genes (Basel),* 11.
SHEPPARD, S. K., DIDELOT, X., MERIC, G., TORRALBO, A., JOLLEY, K. A., KELLY, D. J., BENTLEY, S. D., MAIDEN, M. C., PARKHILL, J. & FALUSH, D. 2013. Genome-wide association study identifies vitamin B5 biosynthesis as a host specificity factor in Campylobacter. *Proceedings of the national academy of sciences,* 110**,** 11923-11927.
SHRIVASTAVA, S. R., SHRIVASTAVA, P. S. & RAMASAMY, J. 2018. World health organization releases global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics. *Journal of Medical Society,* 32**,** 76.
SIGUIER, P., GOURBEYRE, E. & CHANDLER, M. 2014. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiology Reviews,* 38**,** 865-891.
SIMMS, I., FIELD, N., JENKINS, C., CHILDS, T., GILBART, V. L., DALLMAN, T. J., MOOK, P., CROOK, P. D. & HUGHES, G. 2015. Intensified shigellosis epidemic associated with sexual transmission in men who have sex with men--Shigella flexneri and S. sonnei in England, 2004 to end of February 2015. *Euro Surveill,* 20.

SMITH, L. M., FUNG, S., HUNKAPILLER, M. W., HUNKAPILLER, T. J. & HOOD, L. E. 1985. The synthesis of oligonucleotides containing an aliphatic amino group at the 5′ terminus: synthesis of fluorescent DNA primers for use in DNA sequence analysis. *Nucleic acids research,* 13**,** 2399-2412.

SOLAR, G. D., HERNÁNDEZ-ARRIAGA, A. M., GOMIS-RÜTH, F. X., COLL, M. & ESPINOSA, M. 2002. A Genetically Economical Family of Plasmid-Encoded Transcriptional Repressors Involved in Control of Plasmid Copy Number. *Journal of Bacteriology,* 184**,** 4943-4951.

SORBARA, M. T. & PAMER, E. G. 2019. Interbacterial mechanisms of colonization resistance and the strategies pathogens use to overcome them. *Mucosal Immunol,* 12**,** 1-9.

STEENBERGEN, S. M., JIRIK, J. L. & VIMR, E. R. 2009. YjhS (NanS) is required for Escherichia coli to grow on 9-O-acetylated N-acetylneuraminic acid. *Journal of bacteriology,* 191**,** 7134-7139.

STEINEGGER, M. & SÖDING, J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology,* 35**,** 1026-1028.

STICKLAND, H. G., DAVENPORT, P. W., LILLEY, K. S., GRIFFIN, J. L. & WELCH, M. 2010. Mutation of nfxB causes global changes in the physiology and metabolism of Pseudomonas aeruginosa. *Journal of proteome research,* 9**,** 2957-2967.

STUBBENDIECK, R. M. & STRAIGHT, P. D. 2016. Multifaceted Interfaces of Bacterial Competition. *Journal of bacteriology,* 198**,** 2145-2155.

SUN, Q., LAN, R., WANG, J., XIA, S., WANG, Y., WANG, Y., JIN, D., YU, B., KNIREL, Y. A. & XU, J. 2013. Identification and Characterization of a Novel Shigella flexneri Serotype Yv in China. *PLOS ONE,* 8**,** e70238.

SUN, Q., LAN, R., WANG, Y., ZHAO, A., ZHANG, S., WANG, J., WANG, Y., XIA, S., JIN, D., CUI, Z., ZHAO, H., LI, Z., YE, C., ZHANG, S., JING, H. & XU, J. 2011. Development of a multiplex PCR assay targeting O-antigen modification genes for molecular serotyping of Shigella flexneri. *J Clin Microbiol,* 49**,** 3766-70.

SWERDLOW, H. & GESTELAND, R. 1990. Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic acids research,* 18**,** 1415-1419.

SYDOR, A. M., JOST, M., RYAN, K. S., TURO, K. E., DOUGLAS, C. D., DRENNAN, C. L. & ZAMBLE, D. B. 2013. Metal binding properties of Escherichia coli YjiA, a member of the metal homeostasis-associated COG0523 family of GTPases. *Biochemistry,* 52**,** 1788-1801.

TACCONELLI, E., CARRARA, E., SAVOLDI, A., HARBARTH, S., MENDELSON, M., MONNET, D. L., PULCINI, C., KAHLMETER, G., KLUYTMANS, J., CARMELI, Y., OUELLETTE, M., OUTTERSON, K., PATEL, J., CAVALERI, M., COX, E. M., HOUCHENS, C. R., GRAYSON, M. L., HANSEN, P., SINGH, N., THEURETZBACHER, U., MAGRINI, N. & GROUP, W. H. O. P. P. L. W. 2018. Discovery, research, and development of new antibiotics: the WHO priority list of antibiotic-resistant bacteria and tuberculosis. *Lancet Infect Dis,* 18**,** 318-327.

TALUKDER, K. A., ISLAM, M. A., KHAJANCHI, B. K., DUTTA, D. K., ISLAM, Z., SAFA, A., ALAM, K., HOSSAIN, A., NAIR, G. B. & SACK, D. A. 2003. Temporal Shifts in the Dominance of Serotypes of Shigella dysenteriae from 1999 to 2002 in Dhaka, Bangladesh. 41**,** 5053-5058.

TANG, X., CHANG, S., ZHANG, K., LUO, Q., ZHANG, Z., WANG, T., QIAO, W., WANG, C., SHEN, C., ZHANG, Z., ZHU, X., WEI, X., DONG, C., ZHANG, X. & DONG, H. 2021. Structural basis for bacterial lipoprotein relocation by the transporter LolCDE. *Nat Struct Mol Biol,* 28**,** 347-355.

TATUSOVA, T., DICUCCIO, M., BADRETDIN, A., CHETVERNIN, V., NAWROCKI, E. P., ZASLAVSKY, L., LOMSADZE, A., PRUITT, K. D., BORODOVSKY, M. & OSTELL, J. 2016. NCBI prokaryotic genome annotation pipeline. *Nucleic Acids Res,* 44**,** 6614-24.

TEAM, R. C. 2013. R: A language and environment for statistical computing.

TENOVER, F. C. 2006. Mechanisms of Antimicrobial Resistance in Bacteria. *The American Journal of Medicine,* 119**,** S3-S10.

TERRY, C. M., PICKING, W. L., BIRKET, S. E., FLENTIE, K., HOFFMAN, B. M., BARKER, J. R. & PICKING, W. D. 2008. The C-terminus of IpaC is required for effector activities related to Shigella invasion of host cells. *Microb Pathog,* 45**,** 282-9.

THE, H. C., THANH, D. P., HOLT, K. E., THOMSON, N. R. & BAKER, S. 2016. The genomic signatures of Shigella evolution, adaptation and geographical spread. *Nature Reviews Microbiology,* 14**,** 235-250.

THOMPSON, C. N., DUY, P. T. & BAKER, S. 2015. The Rising Dominance of Shigella sonnei: An Intercontinental Shift in the Etiology of Bacillary Dysentery. *PLoS Negl Trop Dis,* 9**,** e0003708.

TOBIN, J. & SCHLEIF, R. 1987a. Positive regulation of the Escherichia coli L-rhamnose operon is mediated by the products of tandemly repeated regulatory genes. *Journal of molecular biology,* 196**,** 789-799.

TOBIN, J. F. & SCHLEIF, R. F. 1987b. Positive regulation of the Escherichia coli L-rhamnose operon is mediated by the products of tandemly repeated regulatory genes. *J Mol Biol,* 196**,** 789-99.

TONKIN-HILL, G., MACALASDAIR, N., RUIS, C., WEIMANN, A., HORESH, G., LEES, J. A., GLADSTONE, R. A., LO, S., BEAUDOIN, C., FLOTO, R. A., FROST, S. D. W., CORANDER, J., BENTLEY, S. D. & PARKHILL, J. 2020. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biology,* 21**,** 180.

TORREZ LAMBERTI, M. F., TERAN, L. C., LOPEZ, F. E., DE LAS MERCEDES PESCARETTI, M. & DELGADO, M. A. 2022. Genomic and proteomic characterization of two strains of Shigella flexneri 2 isolated from infants' stool samples in Argentina. *BMC Genomics,* 23**,** 495.

TOTSIKA, M., WELLS, T. J., BELOIN, C., VALLE, J., ALLSOPP, L. P., KING, N. P., GHIGO, J.-M. & SCHEMBRI, M. A. 2012. Molecular Characterization of the EhaG and UpaG Trimeric Autotransporter Proteins from Pathogenic Escherichia coli. *Applied and Environmental Microbiology,* 78**,** 2179-2189.

TROEGER, C., BLACKER, B. F., KHALIL, I. A., RAO, P. C., CAO, S., ZIMSEN, S. R. M., ALBERTSON, S. B., STANAWAY, J. D., DESHPANDE, A., ABEBE, Z., ALVIS-GUZMAN, N., AMARE, A. T., ASGEDOM, S. W., ANTENEH, Z. A., ANTONIO, C. A. T., AREMU, O., ASFAW, E. T., ATEY, T. M., ATIQUE, S., AVOKPAHO, E. F. G. A., AWASTHI, A., AYELE, H. T., BARAC, A., BARRETO, M. L., BASSAT, Q., BELAY, S. A., BENSENOR, I. M., BHUTTA, Z. A., BIJANI, A., BIZUNEH, H., CASTAÑEDA-ORJUELA, C. A., DADI, A. F., DANDONA, L., DANDONA, R., DO, H. P., DUBEY, M., DUBLJANIN, E., EDESSA, D., ENDRIES, A. Y., ESHRATI, B., FARAG, T., FEYISSA, G. T., FOREMAN, K. J., FOROUZANFAR, M. H., FULLMAN, N., GETHING, P. W., GISHU, M. D., GODWIN, W. W., GUGNANI, H. C., GUPTA, R., HAILU, G. B., HASSEN, H. Y., HIBSTU, D. T., ILESANMI, O. S., JONAS, J. B., KAHSAY, A., KANG, G., KASAEIAN, A., KHADER, Y. S., KHALIL, I. A., KHAN, E. A., KHAN, M. A., KHANG, Y.-H., KISSOON, N., KOCHHAR, S., KOTLOFF, K. L., KOYANAGI, A., KUMAR, G. A., MAGDY ABD EL RAZEK, H., MALEKZADEH, R., MALTA, D. C., MEHATA, S., MENDOZA, W., MENGISTU, D. T., MENOTA, B. G., MEZGEBE, H. B., MLASHU, F. W., MURTHY, S., NAIK, G. A., NGUYEN, C. T., NGUYEN, T. H., NINGRUM, D. N. A., OGBO, F. A., OLAGUNJU, A. T., PAUDEL, D., PLATTS-MILLS, J. A., QORBANI, M., RAFAY, A., RAI, R. K., RANA, S. M., RANABHAT, C. L., RASELLA, D., RAY, S. E., REIS, C., RENZAHO, A. M. N., REZAI, M. S., RUHAGO, G. M., SAFIRI, S., SALOMON, J. A., SANABRIA, J. R., 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. *The Lancet Infectious Diseases,* 18**,** 1211-1228.

TUCKER, N. P., D'AUTRÉAUX, B., SPIRO, S. & DIXON, R. 2006. Mechanism of transcriptional regulation by the Escherichia coli nitric oxide sensor NorR. *Biochemical Society Transactions,* 34**,** 191-194.

TURNER, R. J., WEINER, J. H. & TAYLOR, D. E. 1995. The tellurite-resistance determinants tehAtehB and klaAklaBtelB have different biochemical requirements. *Microbiology,* 141**,** 3133-3140.

TURNER, S. A., LUCK, S. N., SAKELLARIS, H., RAJAKUMAR, K. & ADLER, B. 2001. Nested deletions of the SRL pathogenicity island of Shigella flexneri 2a. *J Bacteriol,* 183**,** 5535-43.

UD-DIN, A. I., WAHID, S. U., LATIF, H. A., SHAHNAIJ, M., AKTER, M., AZMI, I. J., HASAN, T. N., AHMED, D., HOSSAIN, M. A. & FARUQUE, A. S. 2013. Changing trends in the prevalence of Shigella species: emergence of multi-drug resistant Shigella sonnei biotype g in Bangladesh. *PloS one,* 8**,** e82601.

UFFELMANN, E., HUANG, Q. Q., MUNUNG, N. S., DE VRIES, J., OKADA, Y., MARTIN, A. R., MARTIN, H. C., LAPPALAINEN, T. & POSTHUMA, D. 2021. Genome-wide association studies. *Nature Reviews Methods Primers,* 1**,** 59.

VAN DIJK, E. L., AUGER, H., JASZCZYSZYN, Y. & THERMES, C. 2014. Ten years of next-generation sequencing technology. *Trends in Genetics,* 30**,** 418-426.

VANHOMMERIG, E., MOONS, P., PIRICI, D., LAMMENS, C., HERNALSTEENS, J.-P., DE GREVE, H., KUMAR-SINGH, S., GOOSSENS, H. & MALHOTRA-KUMAR, S. 2014.

Comparison of Biofilm Formation between Major Clonal Lineages of Methicillin Resistant Staphylococcus aureus. *PLOS ONE,* 9**,** e104561.
VARGAS, M., GASCON, J., JIMENEZ DE ANTA, M. T. & VILA, J. 1999. Prevalence of Shigella enterotoxins 1 and 2 among Shigella strains isolated from patients with traveler's diarrhea. *Journal of clinical microbiology,* 37**,** 3608-3611.
VERGARA-IRIGARAY, M., FOOKES, M. C., THOMSON, N. R. & TANG, C. M. 2014. RNA-seq analysis of the influence of anaerobiosis and FNR on Shigella flexneri. *BMC Genomics,* 15**,** 438.
VERMA, S. C. & MAHADEVAN, S. 2012. The chbG gene of the chitobiose (chb) operon of Escherichia coli encodes a chitooligosaccharide deacetylase. *J Bacteriol,* 194**,** 4959-71.
VISSCHER, PETER M., BROWN, MATTHEW A., MCCARTHY, MARK I. & YANG, J. 2012. Five Years of GWAS Discovery. *The American Journal of Human Genetics,* 90**,** 7-24.
VOGWILL, T. & MACLEAN, R. C. 2015. The genetic basis of the fitness costs of antimicrobial resistance: a meta-analysis approach. *Evolutionary Applications,* 8**,** 284-295.
VON SEIDLEIN, L., KIM, D. R., ALI, M., LEE, H., WANG, X., THIEM, V. D., CANH, D. G., CHAICUMPA, W., AGTINI, M. D. & HOSSAIN, A. 2006. A multicentre study of Shigella diarrhoea in six Asian countries: disease burden, clinical manifestations, and microbiology. *PLoS medicine,* 3**,** e353.
WAND, M. E., BAKER, K. S., BENTHALL, G., MCGREGOR, H., MCCOWEN, J. W. I., DEHEER-GRAHAM, A. & SUTTON, J. M. 2015a. Characterization of Pre-Antibiotic Era Klebsiella pneumoniae Isolates with Respect to Antibiotic/Disinfectant Susceptibility and Virulence in Galleria mellonella. 59**,** 3966-3972.
WAND, M. E., BAKER, K. S., BENTHALL, G., MCGREGOR, H., MCCOWEN, J. W. I., DEHEER-GRAHAM, A. & SUTTON, J. M. 2015b. Characterization of Pre-Antibiotic Era Klebsiella pneumoniae Isolates with Respect to Antibiotic/Disinfectant Susceptibility and Virulence in Galleria mellonella. 59**,** 3966-3972.
WANG, Z., FAN, G., HRYC, C. F., BLAZA, J. N., SERYSHEVA, I. I., SCHMID, M. F., CHIU, W., LUISI, B. F. & DU, D. 2017. An allosteric transport mechanism for the AcrAB-TolC multidrug efflux pump. *eLife,* 6**,** e24905.
WATSON, R., ROWSOME, W., TSAO, J. & VISENTIN, L. P. 1981. Identification and characterization of Col plasmids from classical colicin E-producing strains. *J Bacteriol,* 147**,** 569-77.
WEI, Y. & MURPHY, E. R. 2016. Shigella Iron Acquisition Systems and their Regulation. *Frontiers in cellular and infection microbiology,* 6**,** 18-18.

WHELAN, F. J., RUSILOWICZ, M. & MCINERNEY, J. O. 2020. Coinfinder: detecting significant associations and dissociations in pangenomes. *Microbial genomics,* 6**,** e000338.

WHITMORE, L. & WALLACE, B. A. 2008. Protein secondary structure analyses from circular dichroism spectroscopy: methods and reference databases. *Biopolymers: Original Research on Biomolecules,* 89**,** 392-400.

WICK, R. R., JUDD, L. M., GORRIE, C. L. & HOLT, K. E. 2017a. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLOS Computational Biology,* 13**,** e1005595.

WICK, R. R., JUDD, L. M., GORRIE, C. L. & HOLT, K. E. 2017b. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol,* 13**,** e1005595.

WILLIAMS, P. C. M. & BERKLEY, J. A. 2018. Guidelines for the treatment of dysentery (shigellosis): a systematic review of the evidence. *Paediatrics and international child health,* 38**,** S50-S65.

WU, Q., SABOKROO, N., WANG, Y., HASHEMIAN, M., KARAMOLLAHI, S. & KOUHSARI, E. 2021. Systematic review and meta-analysis of the epidemiology of vancomycin-resistance Staphylococcus aureus isolates. *Antimicrobial Resistance & Infection Control,* 10**,** 101.

WYCKOFF, E. E., BOULETTE, M. L. & PAYNE, S. M. 2009. Genetics and environmental regulation of Shigella iron transport systems. *Biometals,* 22**,** 43-51.

XU, Y. & ZHOU, N. Y. 2020. MhpA Is a Hydroxylase Catalyzing the Initial Reaction of 3-(3-Hydroxyphenyl)Propionate Catabolism in Escherichia coli K-12. *Appl Environ Microbiol,* 86.

YANG, F., YANG, J., ZHANG, X., CHEN, L., JIANG, Y., YAN, Y., TANG, X., WANG, J., XIONG, Z., DONG, J., XUE, Y., ZHU, Y., XU, X., SUN, L., CHEN, S., NIE, H., PENG, J., XU, J., WANG, Y., YUAN, Z., WEN, Y., YAO, Z., SHEN, Y., QIANG, B., HOU, Y., YU, J. & JIN, Q. 2005. Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic acids research,* 33**,** 6445-6458.

YE, C., LAN, R., XIA, S., ZHANG, J., SUN, Q., ZHANG, S., JING, H., WANG, L., LI, Z. & ZHOU, Z. 2010. Emergence of a new multidrug-resistant serotype X variant in an epidemic clone of Shigella flexneri. *Journal of clinical microbiology,* 48**,** 419-426.

ZAGHLOUL, L., TANG, C., CHIN, H. Y., BEK, E. J., LAN, R. & TANAKA, M. M. 2007. The distribution of insertion sequences in the genome of Shigella flexneri strain 2457T. *FEMS Microbiol Lett,* 277**,** 197-204.

ZAYET, S., KLOPFENSTEIN, T., PIERRON, A., ROYER, P. Y., TOKO, L., GARNIER, P. & GENDRIN, V. 2021. Shigella sonnei, an emerging multidrug-resistant sexually transmitted pathogen in Franche-Comte, France. *Emerg Microbes Infect,* 10**,** 1702-1705.

ZERBINO, D. R. 2010. Using the Velvet de novo assembler for short-read sequencing technologies. *Current protocols in bioinformatics,* 31**,** 11.5. 1-11.5. 12.

ZHANG, Y., LIAO, Y.-T., SALVADOR, A., SUN, X. & WU, V. C. H. 2020. Prediction, Diversity, and Genomic Analysis of Temperate Phages Induced From Shiga Toxin-Producing Escherichia coli Strains. *Frontiers in Microbiology,* 10.

ZOUED, A., BRUNET, Y. R., DURAND, E., ASCHTGEN, M. S., LOGGER, L., DOUZI, B., JOURNET, L., CAMBILLAU, C. & CASCALES, E. 2014. Architecture and assembly of the Type VI secretion system. *Biochim Biophys Acta,* 1843**,** 1664-73.

ZURAWSKI, D. V., MUMY, K. L., FAHERTY, C. S., MCCORMICK, B. A. & MAURELLI, A. T. 2009. Shigella flexneri type III secretion system effectors OspB and OspF target the nucleus to downregulate the host inflammatory response via interactions with retinoblastoma protein. *Molecular microbiology,* 71**,** 350-368.