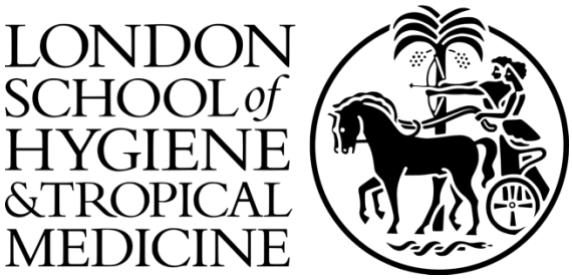


Using zebrafish to study the evolution and pathogenesis of *Shigella*

SYDNEY-LEIGH MILES

Thesis submitted in accordance with the requirements for the degree of

Doctor of Philosophy
of the
University of London

March 2025

Department of Infection Biology
Faculty of Infectious and Tropical Disease
London School of Hygiene & Tropical Medicine

Funded by the BBSRC, London Interdisciplinary Doctoral Programme (LIDo DTP)

Research group affiliations:

Prof. Serge Mostowy (Primary supervisor)

Prof. Kathryn Holt (Secondary supervisor)

Declaration

I, Sydney-Leigh Miles, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Date 26.09.2024

Acknowledgements

First of all, I would like to thank the BBSRC and the London Interdisciplinary Doctoral Training Programme for funding this PhD and making this work possible.

I am hugely grateful to my primary supervisor, Serge Mostowy, for welcoming me into the Mostowy Lab, fresh from my undergraduate studies in the midst of COVID-19. Since then, your constant encouragement, enthusiasm for results at any time of the day and the opportunities you have provided have made my PhD experience incredibly fulfilling. To my secondary supervisor, Kathryn Holt, thank you for sharing a wealth of knowledge on *Shigella* genomics with me and for generally being an outstanding role model over the past three years.

I would like to thank all the fantastic collaborators who made this project possible: Claire Jenkins, Kate Baker, Steve Baker and François Xavier-Weill for sharing clinical isolates; Dilys Santillo and Vanessa Sancho-Shimizu at Imperial College London for help with neutrophil experiments, and Charlotte Chong and Malaka de Silva at the University of Liverpool for help with sequencing. I would also like to thank the BSF team at LSHTM for taking great care of the zebrafish and ensuring the constant availability of eggs.

This PhD has been made so much more enriching (both scientifically and personally!) by the constant support provided by members of the Mostowy Lab, past and present. Special thanks to Vincenzo, who took me under his wing as a rotation student and introduced me to the world of zebrafish. To Margarida, Ana and Dominik – the postdoc superstars who have provided a huge amount of technical guidance, moral support and have generally been a constant source of inspiration (and Kit-Kat breaks). To Kathryn for the constant supply of funny tweets and TikToks. I would also like to thank members of the Holt Lab for valuable genomics discussions, especially Zoe, who has been incredibly patient with my hundreds of questions!

A huge thank you to Mollie and Helen who I met at the start of PhD journey, and who I am sure will now be friends for life. Thank you for your constant willingness to get blueberry matchas, perform terribly in pub quizzes and queue all day to see Taylor Swift – all of which

made these past three years so much more enjoyable! And thank you to anyone I have not mentioned by name but has been there throughout this journey and has endured me talking about dysentery…

I would like to thank my family for their constant belief in my abilities as a first-generation university student. To my Dad, who is not here to see me finish my studies, but who I know would be incredibly proud. To my Nan and Grandad who have provided a lifetime of love and support and have always shown an interest in my work. To my brother Jamie, for the boasting to your friends (and generally boosting my ego!). Lastly, thank you to my Mum, who always encouraged me just to try my best, nourished my love for education from a young age and has done everything possible to allow me to get to this point.

Finally, to Oliver, without the tremendous amount of love and support you have provided throughout, I am sure this PhD would not have been possible. Thank you for your endless ability to make me smile, you have kept me going throughout the toughest of moments.

Abstract

Shigellosis (bacillary dysentery) is a leading cause of diarrhoeal deaths globally, caused by lineages of *Escherichia coli* that convergently evolved to become specialised human pathogens. The independent acquisitions of a large virulence plasmid (pINV) by *Shigella* spp. and enteroinvasive *E. coli* (EIEC) conferred the ability to invade epithelial cells and manifest severe diarrhoeal disease. Within *Shigella* and EIEC, some lineages are significantly more epidemiologically successful than others, but factors underlying their success remain underexplored. This thesis combines genomics, bioinformatic approaches and the use of both *in vivo* and *in cellulo* infection models to explore the pathogenicity of different subgroups of *Shigella* to identify factors that contribute to their epidemiological success in humans.

Firstly, I use the recently emerged serotype O96:H19 (and sequence type (ST) 99) EIEC as a model to explore the early stages of evolution. By reconstructing a dated phylogeny of ST99 *E. coli*, distinct pINV-positive and -negative clusters were identified and subsequent zebrafish infection revealed distinct mechanisms of virulence within these two clusters. This work provides novel insights into the keystone role of pINV acquisition in the virulence of a recently emerged clone.

Next I focus on *Shigella sonnei,* a more established but relatively recently diverged *Shigella* subgroup emerging as the dominant agent of shigellosis worldwide. A collection of epidemiologically relevant clinical isolates was curated and underwent whole genome sequencing to generate fully completed genomes; comparative genomics was then performed to characterise *S. sonnei* lineages and identify lineage-dependent genomic variation. This analysis revealed ongoing adaptive evolution within *S. sonnei*, characterised by the accumulation of insertion sequences, pseudogenisation and structural rearrangements.

Finally, I used a wet-lab approach to characterise different lineages of *S. sonnei in vivo* (using zebrafish) and *in cellulo* (using HeLa cells and human neutrophils), and then experimentally explore factors that contribute to their virulence. Here, the overall aim is to uncover functional

variations contributing to differences in *S. sonnei* pathogenicity and epidemiological success. Overall, this interdisciplinary approach links the epidemiological landscape of *Shigella* to both genome content and pathogenicity *in vivo* for the first time, providing an innovative pipeline to study the evolution of bacterial pathogens and uncover signatures of pathogen success.

List of publications

Below is a list of publications that I have contributed to throughout the duration of this PhD.

MILES, S. L., HOLT, K. E. & MOSTOWY, S. 2024. Recent advances in modelling *Shigella* infection. *Trends in Microbiology*, 32, 917-924.

BROKATZKY, D., GOMES, M. C., ROBERTIN, S., ALBINO, C., **MILES, S. L**. & MOSTOWY, S. 2024. Septins promote macrophage pyroptosis by regulating gasdermin D cleavage and ninjurin-1-mediated plasma membrane rupture. *Cell Chemical Biology*, 31, 1518-1528.e6.

BAKER, K. S.*, HAWKEY, J.*, INGLE, D.*, **MILES, S. L.*** & THE, H. C.* 2024. Chapter 12 - The phylogenomics of *Shigella* spp. In: MOKROUSOV, I. & SHITIKOV, E. (eds.) Phylogenomics. *Academic Press*. * Equal contribution.

MILES, S. L., TORRACA, V., DYSON, Z. A., LÓPEZ-JIMÉNEZ, A. T., FOSTER-NYARKO, E., LOBATO-MÁRQUEZ, D., JENKINS, C., HOLT, K. E. & MOSTOWY, S. 2023. Acquisition of a large virulence plasmid (pINV) promoted temperature-dependent virulence and global dispersal of O96:H19 enteroinvasive *Escherichia coli*. *mBio*, 14, e00882-23.

TORRACA, V., BROKATZKY, D., **MILES, S. L.**, CHONG, C. E., DE SILVA, P. M., BAKER, S., JENKINS, C., HOLT, K. E., BAKER, K. S. & MOSTOWY, S. 2023. *Shigella* Serotypes Associated With Carriage in Humans Establish Persistent Infection in Zebrafish. *The Journal of Infectious Diseases*, 228, 1108-1118.

HUAN, Y. W., TORRACA, V., BROWN, R., FA-ARUN, J., **MILES, S. L**., OYARZÚN, D. A., MOSTOWY, S. & WANG, B. 2023. P1 Bacteriophage-Enabled Delivery of CRISPR-Cas9 Antimicrobial Activity Against *Shigella flexneri*. *ACS Synthetic Biology*, 12, 709-721.

Table of contents

List of figures

List of tables

List of abbreviations

AMR – Antimicrobial resistance

BP – base pair

BSA – Bovine serum albumin

CDSs – Coding sequences

CFU – Colony forming units

CHIM(s) – Controlled human infection model(s)

CR – Congo Red

DIC – Deviance information criteria

DNA – Deoxyribonucleic acid

DPF – Days post fertilisation

EHEC – Enterohaemorrhagic *Escherichia coli*

EIEC – Enteroinvasive *Escherichia coli*

EPEC – Enteropathogenic *Escherichia coli*

ETEC – Enterotoxigenic *Escherichia coli*

G4C – Group 4 capsule

GDP – Gross domestic product

GFP – Green fluorescent protein

H-NS – Histone-like nucleoid-structuring protein

HBV – Hindbrain ventricle

HGCs – Homologous gene clusters

HIC – High-income country

HPF – Hours post fertilisation

HPI – Hours post infection

IS(s) – Insertion sequence(s)

Kbp – Kilobase pair

LCBs – Linear colocalised blocks

LMIC – Low- and middle-income country

LPS – Lipopolysaccharide

M-cell – Microfold cell

Mbp – Megabase pair

MCMC – Markov chain Monte Carlo

MLST – Multilocus sequence type

MOI – Multiplicity of infection

MRCA – Most recent common ancestor

mRNA – Messenger RNA

MSM – Men who have sex with men

OD – Optical density

ONT – Oxford Nanopore Technologies

OUCRU – Oxford University Clinical Research Unit

PAI – Pathogenicity island

PCR – Polymerase chain reaction

PGAP – Prokaryotic genome annotation pipeline

pINV – Virulence plasmid (of *Shigella*)

PVP - Polyvinylpyrrolidone

QRDR – Quinolone resistance determining region

RBC – Red blood cells

RNA – Ribonucleic acid

RPM – Rotations per minute

SEM – Standard error of the mean

SNP – Single nucleotide polymorphisms

SNV – Single nucleotide variations

SRL – *Shigella* resistance loci

ST – Sequence Type

STEC – Shiga toxin producing *Escherichia coli*

T3SS – Type three secretion system

T6SS – Type six secretion system

TSA – Trypticase soy agar

TSB – Trypticase soy broth

UKHSA – United Kingdom Health and Security Agency

UoL – University of Liverpool

UPEC – Uropathogenic *Escherichia coli*

VAGs – Virulence associated genes

WASH – Water, sanitation and hygiene

WT – Wild type

XDR – Extensively drug resistant

Chapter 1. Introduction

1.1. *Shigella* and *Escherichia coli*

Shigella is a genus comprising of Gram-negative bacterial pathogens belonging to the Enterobacteriaceae family, which cause shigellosis (also known as bacterial dysentery). Despite the nomenclature, *Shigella* does not represent a true genus, but instead human-adapted lineages of *Escherichia coli* which were historically separated based on differences in their biochemical properties and pathogenesis (van den Beld and Reubsaet, 2012). Within *Shigella*, there are four recognised serogroups that are separated based on their antigenic properties: *Shigella boydii*, *Shigella dysenteriae, Shigella flexneri* and *Shigella sonnei.* Each subgroup can be further divided into serotypes (19 for *S. boydii*, 19 for *S. dysenteriae* and 15 for *S. flexneri*), with the exception of *S. sonnei* which is comprised of just a single serotype (Karaolis et al., 1994). Complicating the nomenclature further is enteroinvasive *E. coli* (EIEC), a pathotype of diarrheagenic *E. coli* which causes the same clinical symptoms as *Shigella.* EIEC shares many genotypic and phenotypic properties with *Shigella* but does not currently fall within the classical nomenclatural system due to historical differences in pathogenicity (van den Beld and Reubsaet, 2012).

1.2. Evolution and emergence of *Shigella*

The close relationship between *Shigella* and *E. coli* has long been recognised (Luria and Burrous, 1957) and it is now clear from the sequencing of house-keeping genes and more recently, whole genomes, that *Shigella* is interspersed within the broader *E. coli* phylogeny (Pupo et al., 2000, Sims and Kim, 2011, Pupo et al., 1997) (Fig 1.1). The dispersal of *Shigella* and EIEC over multiple clusters of *E. coli* indicates that each 'species', and in some cases, serotypes, have arisen from different *E. coli* ancestors on several independent occasions (Pupo et al., 2000).

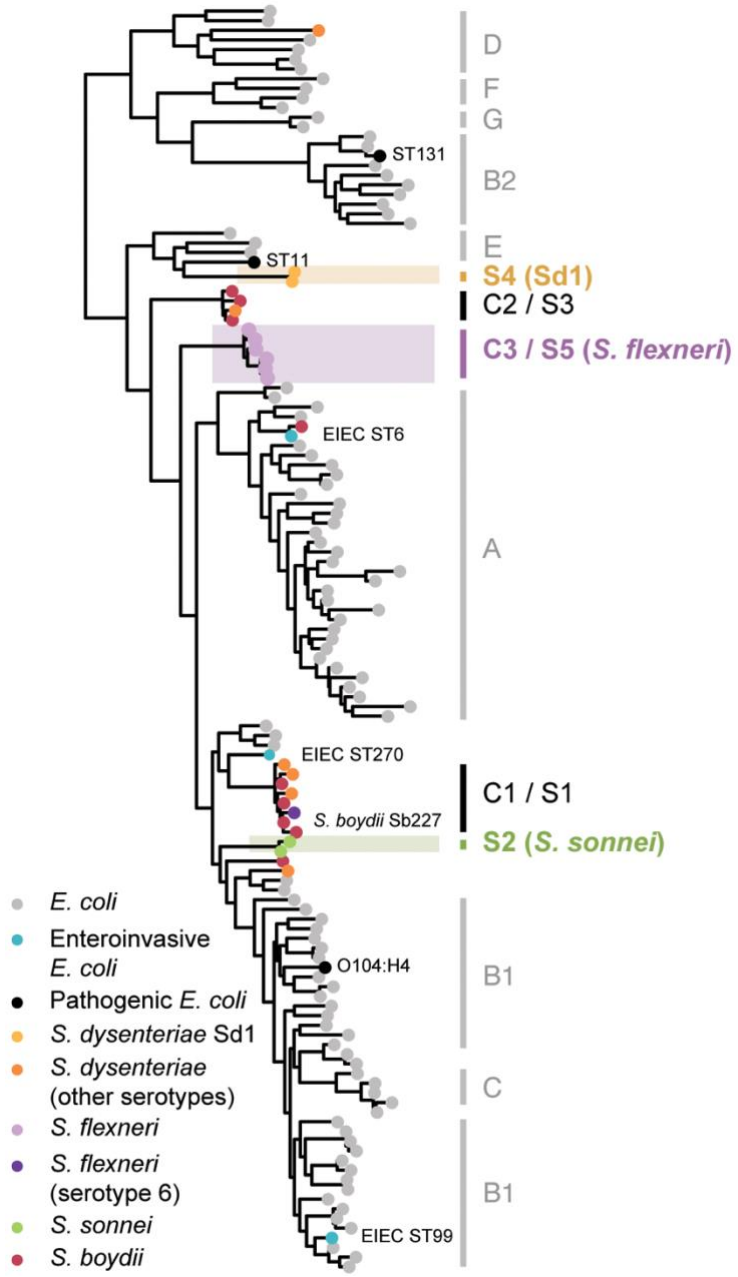


Figure 2.1. Midpoint rooted maximum likelihood core gene phylogeny of *Escherichia coli* and *Shigella* genomes.

Highlighted in colour are the different subgroups of *Shigella* and enteroinvasive *E. coli* (EIEC), with *E. coli* nodes shown in grey. *E. coli* phylogroups are labelled in grey. *Shigella* subgroups are interspersed throughout the phylogeny, among different *E. coli* phylogroups, indicating the existence of multiple *E. coli* ancestors. Figure taken from Hawkey et al., 2020.

A defining event in the emergence of *Shigella* and EIEC is the acquisition of a ~210-240 kilobase pair (kbp) plasmid (pINV) which encodes an arsenal of virulence factors, including a type three secretion system (T3SS), essential for invading the human gut epithelium (Schroeder and Hilbi, 2008). Genetic analysis of pINV has revealed the presence of two distinct forms (pINV A and pINV B), which have different plasmid incompatibility groupings and are interspersed across the phylogeny (Lan et al., 2001), further supporting the notion that *Shigella* and EIEC have arisen by convergent evolution. Aside from pINV acquisition, several other gain- and loss-of-function changes (which may occur in various combinations in different subgroups) have been established as stepwise evolutionary changes in the transition to becoming a human-adapted pathogen (Fig 1.2).

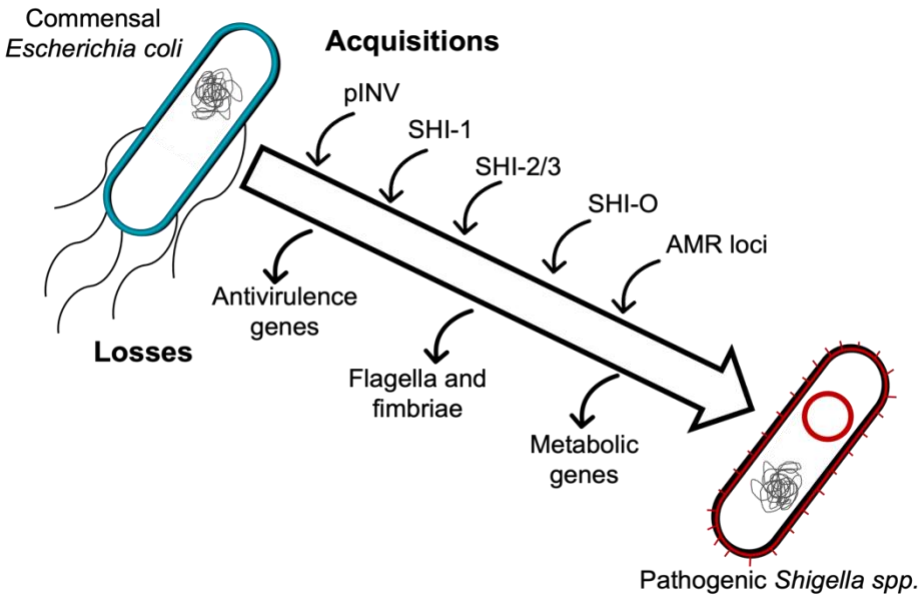


Figure 1.2. The proposed evolutionary transition from commensal *Escherichia coli* towards pathogenicity in *Shigella*. The formation of *Shigella* pathovariants likely begins with the acquisition of large invasion plasmid, pINV, which confers the ability to invade human epithelial cells. Following pINV acquisition, a series of stepwise changes occur whereby virulence and resistance determinants are gained, and genes that interfere with pathogenicity (antivirulence genes and immunogenic flagella and fimbriae), as well as genes that are no longer needed (like metabolic genes) are lost. SHI = *Shigella* pathogenicity island. AMR = antimicrobial resistance. Figure adapted from Baker et al, 2024.

Shigella has also acquired chromosomally encoded gene clusters which encode for additional virulence factors known as pathogenicity islands (PAIs). *Shigella* island 1 (SHI-1) encodes for autotransporter proteases (Pic and SigA) and an enterotoxin (ShET1) and is present fully in many *S. flexneri* serotype 2a isolates and partially in other *Shigella* subgroups (Vargas et al., 1999). SHI-2 and SHI-3 encode aerobactin synthesis and transport systems and are distributed in *S. flexneri* and *S. boydii* respectively (Vokes et al., 1999, Purdy and Payne, 2001, Moss et al., 1999). SHI-O encodes genes which enable lipopolysaccharide (LPS) modification, resulting in serotype switching and immune escape in *S. flexneri* (Guan et al., 1999). Finally, the acquisition of genetic material conferring antimicrobial resistance has been fundamental to the success and dispersal of *Shigella*. The *Shigella* resistance loci (SRL) is a cluster of chromosomally encoded antimicrobial resistance determinants which confers resistance to ampicillin, streptomycin, chloramphenicol and tetracycline and is widespread among clinical isolates of all *Shigella* subgroups (Turner et al., 2003). Acquisitions of these PAIs, either partially or in full, have also been documented throughout different EIEC lineages, denoting a further example of convergent evolution.

As is common for many specialised human pathogens, the process of reductive evolution through gene loss and inactivation is observed in *Shigella*. Genome degradation has been well documented in many important human pathogens: *Bordetella pertussis* (Parkhill et al., 2003)*, Mycobacterium tuberculosis* (Stinear et al., 2008), *Salmonella* (Langridge et al., 2015, Pulford et al., 2021) and *Rickettsia* (Diop et al., 2018) and is associated with an increase in pathogenicity in many cases (Murray et al., 2021). Following the colonisation of a new host, a strong selection pressure is applied for the pathogen to adapt, resulting in genome optimisation. This includes the loss of genes no longer necessary to survive such as catabolic genes (where the bacterium can now scavenge metabolites from a new nutrient-rich environment) and the loss of genes that might interfere with the newfound pathogenic lifestyle (Ochman and Moran, 2001). In *Shigella,* gene deletion or inactivation is mostly mediated by

the dynamics of insertion sequences (ISs), small mobile elements that can mobilise within the genome, disrupt coding sequences and mediate genome rearrangements (Zaghloul et al., 2007, Yang et al., 2005). Analysis of IS dynamics across *Shigella* and *E. coli* revealed that *Shigella* and EIEC contain significantly more copies of ISs in contrast to other pathotypes of diarrheagenic *E. coli* and non-pathogenic *E. coli* (Hawkey et al., 2020).

Some genes which have been convergently lost, known as antivirulence genes, can interfere with pathogenicity if reintroduced (Bliven and Maurelli, 2012). In *Shigella* and EIEC, several antivirulence genes have been described: *nadA/B*, which inhibits cellular invasion when expressed (Prunier et al., 2007); *cadA*, which blocks the activity of enterotoxin *senA* (Casalino et al., 2003)*; speG*, which inhibits tolerance of oxidative stress (Barbagallo et al., 2011) when expressed and *ompT,* which interferes with actin-based motility by cleaving IcsA (Maurelli et al., 1998). A genetic study revealed that *cadA* loss is mediated by distinct genomic rearrangements within the different *Shigella* subgroups (Day et al., 2001), highlighting the convergent nature of gene loss in *Shigella* and its clear importance in the transition towards pathogenicity. As well as antivirulence genes, the loss of antigenic components such as flagella and fimbriae (Al Mamun et al., 1997, Bravo et al., 2015) has been documented throughout *Shigella* and EIEC, likely conferring reduced immunogenicity and facilitating host immune evasion. Overall, driven by a series of gene acquisitions and purifying selection, *Shigella* has evolved on multiple independent occasions to efficiently infect humans without compromising its fitness.

1.3. Shigellosis and its epidemiology

The ingestion of as few as 10-100 *Shigella* cells is enough to establish shigellosis, which develops after an incubation period of 1-3 days (DuPont et al., 1989). Transmission usually occurs through consumption of contaminated food or water or through direct contact with an infected person. Shigellosis can manifest in a wide range of clinical symptoms, but classical symptoms include bloody and mucoid diarrhoea, abdominal cramps and fever (Kotloff et al., 2018). In healthy, immunocompetent adults with access to adequate hydration and nutrition,

shigellosis is typically self-limiting and antibiotic treatment is recommended only in severe or complicated cases. Despite this, shigellosis is still the second leading cause of diarrhoeal deaths, causing ~216,000 deaths each year, the burden of which disproportionately affects children under 5 years old in low- and middle-income countries (LMIC) (Khalil et al., 2019). In this demographic, deaths, as well as stunted growth and neurological defects, are associated with malnutrition and lack of access to water, sanitation, and hygiene (WASH) and treatment interventions (Khalil et al., 2018). In high-income countries (HIC), the demographic is different, with infections linked to returning international travellers, or as an endemic sexually transmitted infection particularly affecting men who have sex with men (MSM) (Charles et al., 2022).

When Kyoshi Shiga first described *Shigella* in 1898, *S. dysenteriae* was responsible for most cases of shigellosis but today, along with *S. boydii,* they make up only ~5% of cases (The et al., 2016). Most contemporary infections are instead caused by *S. flexneri* and *S. sonnei*, with *S. flexneri* historically considered to be most prevalent in LMIC and *S. sonnei* accounting for most infections in HIC (Kotloff et al., 1999). However, an intercontinental change in epidemiology has been noted, with *S. sonnei* overtaking *S. flexneri* in many countries undergoing economic transition (Thompson et al., 2015). This phenomenon has been observed in several countries in Asia and South America (Fullá et al., 2005, Sousa et al., 2013, Qu et al., 2012, Vinh et al., 2009), with an increase in gross domestic product (GDP) tightly linked to an increase in the proportion of infections caused by *S. sonnei* (Ram et al., 2008).

The exact explanation behind the shift in epidemiology is unclear but several theories have been put forward, with general hints at discrete responses to environmental pressures between the two subgroups. Firstly, it has been proposed that exposure to waterborne pathogen *Plesiomonas shigelloides* (which shares an identical O-antigen structure with *S. sonnei* (Shepherd James et al., 2000)) may provide human populations with passive immunisation against *S. sonnei.* Passive immunisation is then likely to be depleted upon the

introduction of better access to WASH, ultimately allowing *S. sonnei* to thrive in infecting a newly naïve population (Sack et al., 1994).

Another hypothesis proposed is that *S. sonnei* may be able to use environmental amoeba as a protective niche, like *Legionella pneumophila,* allowing it to circumvent the bactericidal effects of water sanitation. An initial study suggested that *S. dysenteriae* and *S. sonnei* could survive in *Acanthamoeba castellani* for an extended period of time (Saeed et al., 2008), but a more recent study directly comparing *S. flexneri* and *S. sonnei* survival within *A. castellani* challenged this, finding that both subgroups were quickly degraded and neither could replicate intracellularly (Watson et al., 2018).

It has also been suggested that *S. sonnei* may have an innate competitive advantage over *S. flexneri* since it is more recently diverged from, and is genetically closer to *E. coli* (Thompson et al., 2015). In this case, genes retained from an *E. coli* ancestor may enable *S. sonnei* to occupy a wider range of niches, and thus increase its chances of extracellular survival, as compared to *S. flexneri* which is considered to be more adapted to a human host (Hawkey et al., 2020, Hershberg et al., 2007). In the same way, it has been proposed that *S. sonnei* may also be more adept at acquiring advantageous genetic material (such as antimicrobial resistance determinants) from other *E. coli* due to differences in restriction systems (Thompson et al., 2015); a recent study additionally demonstrated that *S. sonnei* could acquire extensively drug-resistant (XDR) plasmids from commensal *E. coli* (Thanh Duy et al., 2020).

Finally, *S. sonnei* has been shown to harbour genes which encode for a type six secretion system (T6SS) (Anderson et al., 2017), although the functionality has recently been questioned (De Silva et al., 2023a). Many contemporary isolates also encode colicins, small secreted bacteriocin molecules which efficiently kill sensitive bacterial strains (Leung et al., 2024). The abundance of interbacterial competition weapons in *S. sonnei*, but not *S. flexneri*, could confer a competitive edge, whereby *S. sonnei* not only has greater access to resources

and space but may also colonise a host more efficiently by eliminating protective microbiota (Anderson et al., 2017).

1.4. Pathogenesis of *Shigella*

Most of what is known about *Shigella* pathogenesis is derived from *S. flexneri* infection, which has become a paradigm for studying host-pathogen interactions in cellular microbiology. The defining event of *Shigella* infection is the bacterial invasion of epithelial cells in the human gut and colonic mucosa, which is facilitated by the virulence factors encoded on pINV (Sansonetti et al., 1982). Within pINV, a ~31 kbp region termed the 'entry region' is highly conserved between *Shigella* and EIEC and encodes the T3SS and its effector proteins (Fig 1.3). The T3SS acts as a syringe-like apparatus, delivering effector proteins to a host cell which aid in cellular invasion, modulation of the host immune response and cell death processes (Parsot, 2009).

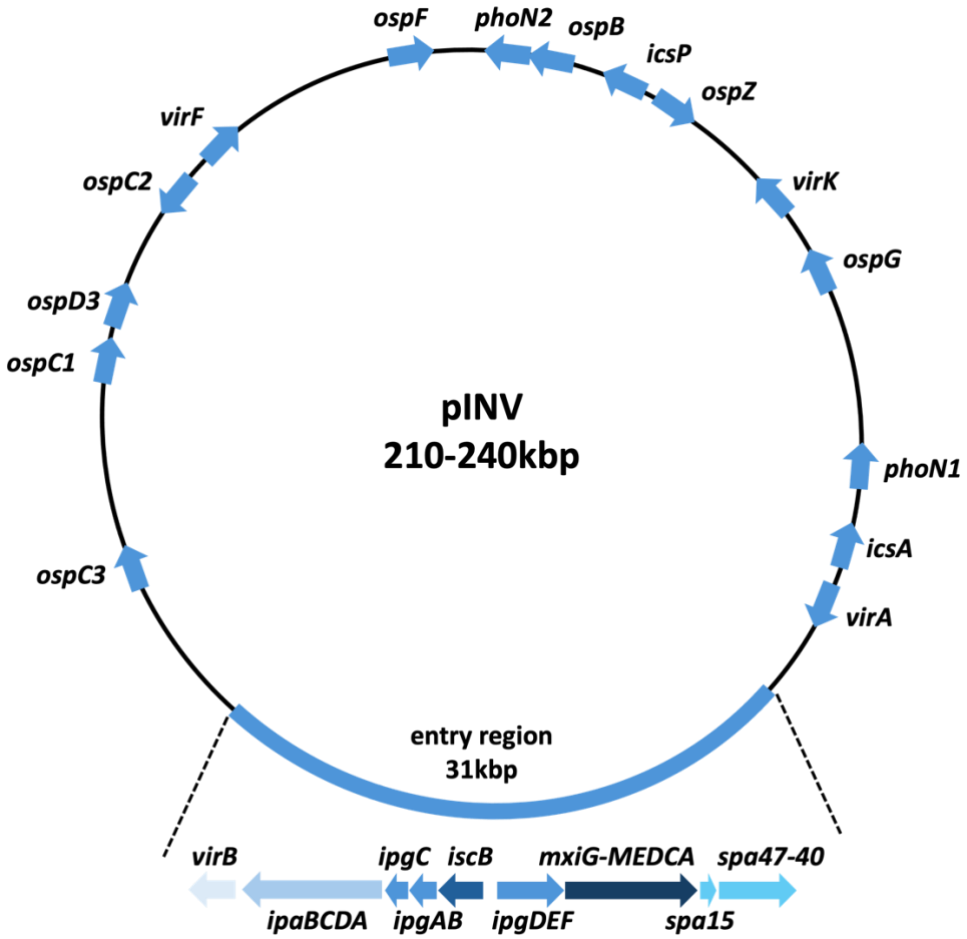


Figure 1.3. Schematic of the invasion plasmid (pINV) in *Shigella* and EIEC. Genetic organisation of characterised virulence genes shown in blue (not to scale). The entry region is conserved between *Shigella* subgroups and encodes the components necessary for type three secretion system (T3SS) assembly. Kbp = kilobase pair. Adapted from Pasqua et al, 2017.

The T3SS is the primary driver of virulence in most *Shigella* subgroups (apart from *S. dysenteriae,* which also harbours the cytotoxic Shiga toxin), and its activity is tightly regulated, only activated when bacteria sense the correct environmental cues, which include pH, temperature and oxygen availability (Nakayama and Watanabe, 1995, Maurelli and Sansonetti, 1988, Marteyn et al., 2010). Following ingestion, *Shigella* must overcome the acidic environment of the human digestion system, a process aided by the modulation of

surface polysaccharides, such as lipopolysaccharide (LPS) (Martinić et al., 2011) or capsular polysaccharide (Caboni et al., 2015). Orchestration of the virulence factors is under the control of a regulatory cascade, with *virF* and *virB* acting as key regulators (Prosseda et al., 2004, Di Martino et al., 2016, Tobe et al., 1993). As an increase in temperature associated with ingestion by a host (> 30 °C) is detected, *virF* expression is induced and the VirF protein binds upstream of the *virB* promoter region, counteracting the silencing effect of histone-like nucleoid-structuring protein (H-NS). Once *virB* is activated, it can then activate the genes required for T3SS assembly (the *mxi-spa* locus) and effector production (Schroeder and Hilbi, 2008).

Shigella must then adhere to its target host cells; interestingly, *Shigella* lacks classical adhesins, but pleiotropic surface protein IcsA has been shown to have adhesive properties which facilitate this interaction (Brotcke Zumsteg et al., 2014). Once direct contact is made with a host microfold cell (M cell), the T3SS is used to secrete bacterial effector proteins into the cell, beginning the process of internalisation. Host actin remodelling is initiated by the Ipg and Ipa effectors, creating membrane ruffles, which promote cell entry (Hachani et al., 2008). Once intracellular, *Shigella* is contained inside a phagocytic vacuole from which it can rapidly escape, with the aid of membrane-lysing effector protein IpaB, to gain access to the cytosol (High et al., 1992). *Shigella* can then replicate in the cytosol and subvert the host cytoskeleton to polymerise actin tail structures in an IcsA-dependent manner, allowing bacteria to spread to neighbouring cells (Bernardini et al., 1989). As well as spreading from cell to cell, *Shigella* is delivered into the basolateral side and engulfed by resident macrophages. Inside macrophages, rapid bacterial replication and the release of effector proteins cause inflammasome-dependent pyroptosis (a form of regulated cell death), ultimately lysing the host cell and releasing *Shigella* to begin the infectious cycle once more (Fig 1.4) (Zychlinsky et al., 1992, Rojas-Lopez et al., 2023).

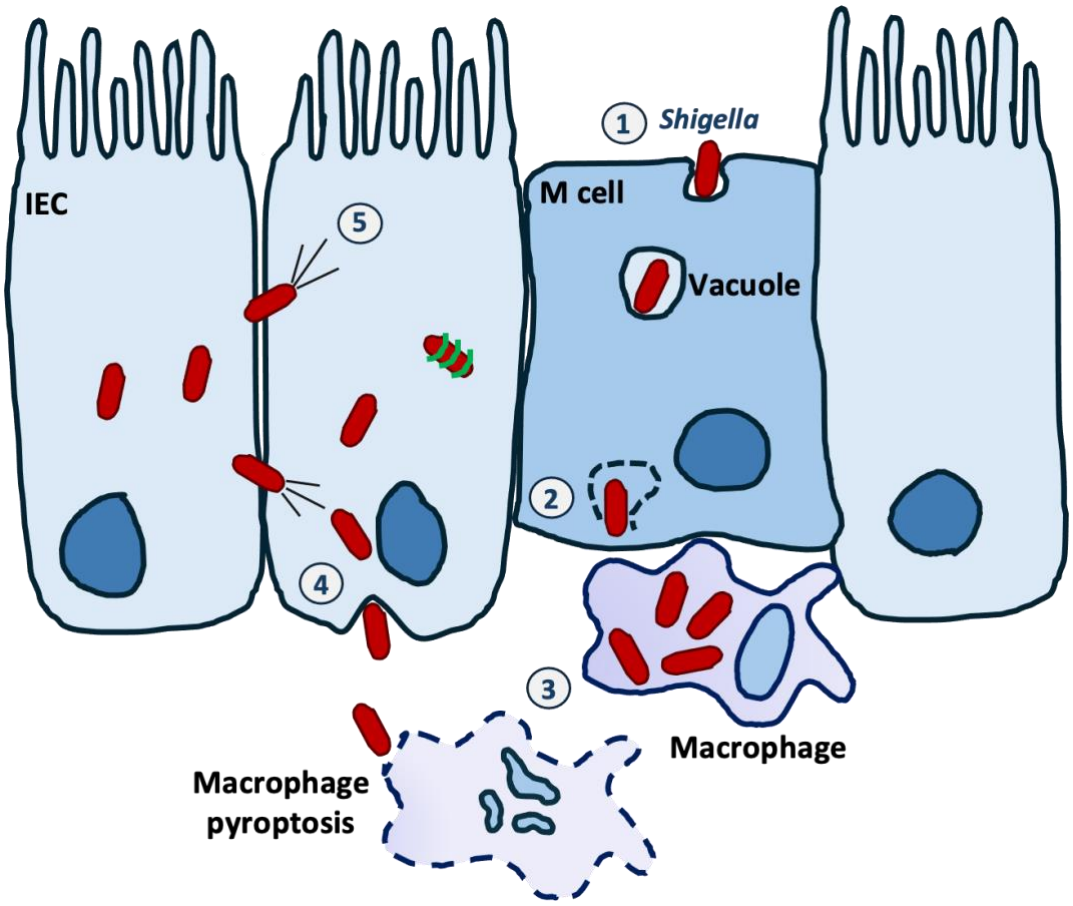


Figure 1.4. Schematic representation of *Shigella* pathogenesis. 1. Contact with a host cell activates the type three secretion system (T3SS) of *Shigella,* stimulating membrane remodelling and *Shigella* uptake; **2.** *Shigella* escapes its phagocytic vacuole and transcytoses through the M cell; **3.** Once on the basolateral side, *Shigella* is engulfed by resident macrophages, where replication and T3SS effector release leads to pyroptosis; **4.** *Shigella* can then infect neighbouring intestinal epithelial cells (IECs) from the basolateral side and replicate cytosolically; **5.** *Shigella* hijacks the host cytoskeleton to generate actin tails allowing cell to cell spread, continuing the infectious cycle.

Since its establishment as a human-adapted pathogen, *Shigella* has been in an evolutionary arms race, evolving elegant mechanisms which aid its evasion of the host immune response and promote virulence. *Shigella* are considered non-motile (having convergently lost the flagellin machinery) but to move intracellularly, *Shigella* can recruit host F-actin to its poles and form actin-tails, a process which is dependent on pINV encoded outer membrane protein, IcsA (Goldberg and Theriot, 1995). During actin polymerisation a complex of host cytoskeletal

proteins, termed septins, can recognise *Shigella,* block actin-tail formation and entrap bacteria in cage-like structures, before targeting them to autophagy (Mostowy et al., 2010, Mostowy et al., 2011, Krokowski et al., 2018). Recently, *Shigella* effector protein OspG has been described to counteract this process, by promoting the ubiquitination of septins and targeting them for degradation (Xian et al., 2024). Together, the processes of invasion, cell to cell spread and cell death induce a massive inflammatory response driven by nuclear factor kappa B (NF-kB) signalling and neutrophil recruitment which damage the epithelium and drive the classical symptoms of dysentery (Ashida et al., 2015).

To survive in the cytosol, *Shigella* must avoid being targeted by cell-autonomous immunity mechanisms (Randow et al., 2013). For example, work has shown that *Shigella* is recognised by autophagy machinery but bacterial effector protein IcsB intervenes with bacterial recognition and destruction, diverting the autophagic process (Mostowy et al., 2011). Additionally, *Shigella* evades killing by pyroptosis through the post-translational modification of caspase-4/11, mediated by the OspC3 effector (Li et al., 2021). In total, there are ~30 effector proteins delivered by the T3SS, each with diverse properties, designed to modulate intracellular mechanisms, subvert the host immune response and facilitate survival as an intracellular pathogen.

1.5. Publication: Recent advances in modelling *Shigella* infection

The contents of this sub-chapter are published as a review article in Trends in Microbiology.

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH2006138	**Title**	Miss
First Name(s)	Sydney-Leigh		
Surname/Family Name	Miles		
Thesis Title	Recent advances in modelling *Shigella* infection		
Primary Supervisor	Serge Mostowy		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

<u>SECTION B – Paper already published</u>

Where was the work published?	Trends in Microbiology		
When was the work published?	September 2024		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	The review was written by me with input and guidance from both supervisors, Prof, Kathryn Holt and Prof. Serge Mostowy.

SECTION E

Student Signature	Sydney-Leigh Miles
Date	04/09/2024

Supervisor Signature	Serge Mostowy
Date	04/09/2024

Trends in Microbiology

CellPress OPEN ACCESS

Review

Recent advances in modelling *Shigella* infection

Sydney L. Miles,[1] Kathryn E. Holt,[1,2] and Serge Mostowy[1,*]

***Shigella* is an important human-adapted pathogen which contributes to a large global burden of diarrhoeal disease. Together with the increasing threat of antimicrobial resistance and lack of an effective vaccine, there is great urgency to identify novel therapeutics and preventatives to combat *Shigella* infection. In this review, we discuss the development of innovative technologies and animal models to study mechanisms underlying *Shigella* infection of humans. We examine recent literature introducing (i) the organ-on-chip model, and its substantial contribution towards understanding the biomechanics of *Shigella* infection, (ii) the zebrafish infection model, which has delivered transformative insights into the epidemiological success of clinical isolates and the innate immune response to *Shigella*, (iii) a pioneering oral mouse model of shigellosis, which has helped to discover new inflammasome biology and protective mechanisms against shigellosis, and (iv) the controlled human infection model, which has been effective in translating basic research into human health impact and assessing suitability of novel vaccine candidates. We consider the recent contributions of each model and discuss where the future of modelling *Shigella* infection lies.**

Highlights

Shigella is an important human-adapted pathogen for which there is an increasing incidence of antimicrobial resistance and no effective vaccine.

Modelling *Shigella* infection of humans has historically been difficult, but new technologies and animal models have emerged to recapitulate key hallmarks of shigellosis.

The use of organ-on-chip technology, zebrafish infection models, transgenic mouse models, and human challenge studies all uniquely contribute to our understanding of *Shigella* infection biology.

Recent advances have illuminated our understanding of *Shigella* pathogenesis, guiding vaccine strategies and moving us closer to human health impact.

Introduction to *Shigella*

Shigella is not a true genus, but is the name given to human-adapted lineages of *Escherichia coli* that cause shigellosis (bacterial dysentery). *Shigella* is further divided into the species *Shigella boydii, Shigella dysenteriae, Shigella flexneri*, and *Shigella sonnei*. *S. flexneri* and *S. sonnei* are **endemic** (see Glossary) in most regions and account for the largest proportion of contemporary infections. By contrast, *S. dysenteriae* expresses a toxin and causes dysentery **epidemics** in crowded settings with poor sanitation and hygiene, whilst *S. boydii* is rare and mainly found in South Asia [1]. Shigellosis is characterised by bacterial invasion of the human gut epithelium, which can result in a range of clinical manifestations, including watery or bloody diarrhoea, fever, and abdominal pain [2]. Shigellosis, although typically self-limiting, still causes ~216 000 deaths each year [1], with the greatest burden of disease falling on children under 5 years old in low- and middle-income countries (LMICs) where mortality is associated with malnutrition [3,4]. *Shigella* can spread through ingestion of contaminated food and water, and as such, disease incidence is associated with limited access to water, sanitation, and hygiene (WASH). However, *S. flexneri* and *S. sonnei* can also be sexually transmitted, and in high-income countries (HICs) with strong WASH infrastructure, shigellosis is increasingly linked either to travel to regions where *Shigella* is endemic, or the sustained sexual transmission between men who have sex with men (MSM) [5]. The transmission of *Shigella* within the MSM community has been associated with the clonal expansion of multidrug and extensively drug resistant *Shigella* genotypes [6,7], presenting a major public health issue as treatment options become extremely limited.

Since mice are naturally resistant to shigellosis, alternative models have long been sought to fully understand the determinants underlying *Shigella* pathogenesis (reviewed in [8]). The first animal model of *Shigella* infection was established as early as 1955, when the Sereny test (a guinea pig keratoconjunctivitis model) was used to differentiate between invasive and non-invasive

[1]Department of Infection Biology, London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK
[2]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria 3004, Australia

*Correspondence: Serge.Mostowy@lshtm.ac.uk (S. Mostowy).

isolates [9]. Seminal work carried out using this model revealed that *Shigella* could spontaneously become avirulent, losing the ability to cause keratoconjunctivitis. This work provided the first clues towards understanding that the invasive phenotype of *Shigella* is dependent on the presence of a **virulence plasmid** (**pINV**), which encodes a **type three secretion system** (**T3SS**) [10,11] and is often unstable at environmental temperatures [12]. More recently, the use of guinea pigs has diverged away from the keratoconjunctivitis model, and more physiologically relevant guinea pig infection models (such as the intra-rectal challenge model) have been implemented for protective efficacy studies [13]. A rabbit ileal loop model, in which a ligated section of the ileum (small intestine) from adult rabbits is inoculated with bacteria, was then used to formulate a detailed model of *Shigella* pathogenesis in an intestinal context [14]. Work performed using this model revealed the precise route in which *Shigella* traverses through the intestinal M cells [15], before being engulfed by macrophages, which ultimately succumb to **pyroptosis** [16,17], triggering the characteristic inflammatory hallmarks of *Shigella* infection. Both guinea pigs and rabbits are small mammalian models able to overcome some of the barriers posed when modelling human-adapted infections, and their significant contributions enhanced our understanding of *Shigella* pathogenesis. Both models, however, neglect the natural route of *Shigella* infection and the complexity and costs of maintaining small mammalian models can be prohibitive to their widespread use.

The emergence of extensively drug-resistant *Shigella* infections [6], along with the absence of an effective vaccine, highlights the necessity to develop novel therapeutic and preventative options for *Shigella*. In this review, we discuss novel technologies and alternative animal models to study *Shigella* infection (Figure 1), the unique contributions each model has provided in recent years, and where the future of modelling *Shigella* and other human-adapted infections may lie.

Organ-on-chip reveals the importance of biomechanical forces during *Shigella* infection

The intracellular lifestyle of *Shigella* has led to its establishment as a paradigm for studying host–pathogen interactions (reviewed in [18]), with many important discoveries (including actin-based motility [19,20] and **septin** cage entrapment [21,22]) having come from the use of immortalised monolayers of cells. Whilst the contribution of these cellular models is clear, *Shigella* infects static monolayers of cells at an infamously low efficiency, which may be explained by the lack of

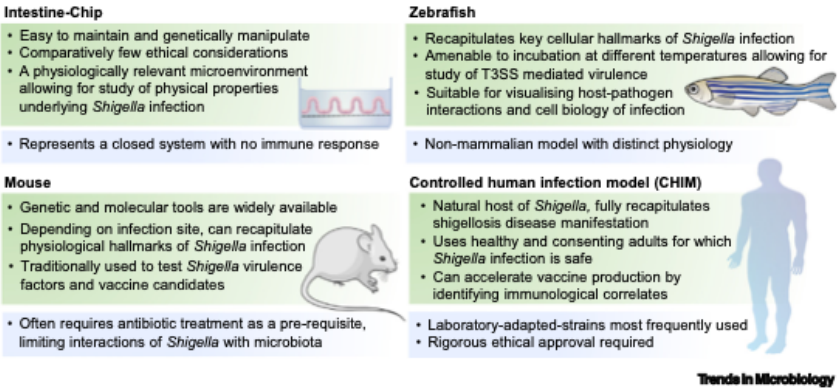


Figure 1. Summary of key advantages and disadvantages of *Shigella* infection models. Advantages of each model in green, and disadvantages of each model in blue. Icons are adapted from those by Servier and DBCLS, both licensed under CC-BY 3.0 Unported. Abbreviation: T3SS, type 3 secretion system.

Glossary

Bacterial persistence: antibiotic-tolerant cells that can establish long-term infection.
Bacteriophage (phage): a type of virus that can infect bacterial cells; phages have been exploited for treatment of multidrug-resistant bacterial infections.
***Bdellovibrio*:** a Gram-negative predatory bacterium that can prey on other bacterial species; it has been used safely for the treatment of multidrug-resistant *Shigella* infection in zebrafish.
Effector protein: a protein secreted by bacteria into host or bacterial cells to manipulate the outcome of host–pathogen or competitive interactions.
EIEC: enteroinvasive *E. coli*; a close relative of *Shigella* that causes bacterial dysentery in the same manner.
Endemic: constant presence of a disease within a particular area.
Epidemic: rapid spread of a disease within a short period of time.
Gasdermin D: a membrane pore-forming protein involved in pyroptosis.
IgA: immunoglobulin A, an antibody found in mucosal membranes.
IgG: immunoglobulin G, an antibody found in blood.
Inflammasomes: multiprotein oligomers that control activation of the inflammatory immune response.
Interferons: signaling proteins released by a host in response to infection; they are involved in controlling *Shigella* infection.
O-antigen: surface polysaccharide of Gram-negative bacteria, a key determinant of *Shigella* serotype-specific immunological protection.
Organoid: an engineered *in vitro* 3D tissue that uses differentiated, self-organising cells to recreate complex microenvironments.
Peristalsis: repeated involuntary contraction of intestinal muscles.
Pyroptosis: a form of caspase-1-dependent programmed cell death that is inflammatory.
Septins: GTP-binding cytoskeletal proteins that can assemble into filaments and cage-like structures, entrapping bacteria for host defence.
Trained innate immunity: training of immune cells via epigenetic modifications to better respond to a secondary infection.
Type three secretion system (T3SS): a needle-like structure facilitating secretion of effector proteins into host cells; it is used for invasion of host cells.

polarisation and differentiation in immortalised epithelial cell lines; however, the development of human **organoid** models with a similar infection rate disputes this hypothesis [23]. Recent advances in organoid and organ-on-chip technology offers promise to overcome these limitations and increase the potential of *in vitro* methodologies. The development of a nutrient-deprived organoid model, and subsequent infection of malnourished organoid-derived cells, showed that *Shigella* infects nutrient-deprived duodenal epithelial cells at a higher rate than healthy monolayers [24]. Increased invasion under nutrient deprived conditions may be reflective of the increased severity of shigellosis seen in LMICs, paving the way for a new and important model for *Shigella* pathogenesis in LMICs. Similarly, Intestine-Chip employs polarised Caco-2 cells (epithelial cells derived from a colon carcinoma) which self-organise into microvilli-like structures to generate a 3D colonic epithelial model; it also applies a cyclic vacuum and continuous flow to recreate cyclic stretching motions resembling **peristalsis** [25]. Intestine-Chip was used to show that *Shigella* is significantly more invasive when exposed to a more physiologically realistic environment and established that *S. flexneri* can exploit the complex organisation of epithelial cells to colonise crypt-like structures of the epithelium which protects against the flow of luminal fluid. This model was then used to further investigate the role of peristalsis on the virulence of *S. flexneri,* and revealed that, in the presence of physical cues, *S. flexneri* presents increased T3SS activation and cell-to-cell spread (Figure 2A), and is able to more efficiently colonise intestinal crypts [26].

Work using the Intestine-Chip model challenged the long-standing hypothesis that *Shigella* invades poorly from the apical cell side, revealing instead, that when mechanical forces mimicking natural intestinal processes are applied, *S. flexneri* can achieve significant invasion of enterocytes directly from the lumen. Whilst the advances made using the Intestine-Chip model are transformative, it is important to consider that it still represents a closed system and ultimately, the incorporation of immune cells will be important to recapitulate shigellosis in a more complex, integrated environment. Overall, lessons learned from modelling *Shigella* infection on a chip have highlighted the significance in mimicking physical cues, which can be challenging in non-native *in vivo* models for human-adapted pathogens like *Shigella*. In the future, organoid and organ-on-chip models may be used to re-visit long-standing hypotheses to update our knowledge of *Shigella*–host interactions and may be used to reduce the number of animals used in experimental modelling.

Type six secretion system (T6SS): a needle-like structure facilitating secretion of effector proteins into neighbouring bacterial cells; it is used in bacterial competition.
Virulence plasmid (pINV): a large plasmid (~220 kbp) encoding virulence factors required for *Shigella* infection.

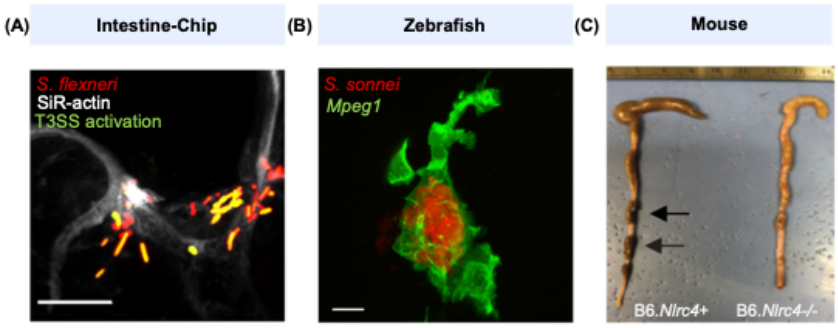


Figure 2. Striking findings from *Shigella* infection models. (A) Fixed image of *Shigella flexneri* (TSAR, a strain which expresses GFP upon T3SS activation) interacting with Intestine-Chip, showing that actin spots frequently colocalise with T3SS activation, highlighting a role for T3SS engagement in cell-to-cell spread. Adapted from [26]. (B) *Shigella sonnei* engulfed inside a zebrafish macrophage *in vivo*, showing that *S. sonnei* can establish persistent infections within macrophages. Adapted from [34]. (C) Representative cecum and colon from B6.*Nlrc4*$^{+}$ and B6.*Nlrc4*$^{-/-}$ mice infected with *S. flexneri*. These results show cecum tissue thickening (resulting in cecum shrinking) and macroscopic oedema in *Nlrc4*-deficient mice. Adapted from [47]. Abbreviation: T3SS, type 3 secretion system.

The versatile *Shigella*–zebrafish infection model

Zebrafish are widely used in the field of infection biology to study host–pathogen interactions with the natural fish pathogen *Mycobacterium marinum* [27]. A decade of work has shown that the *Shigella*–zebrafish model recapitulates key hallmarks of shigellosis (such as macrophage cell death, cell-to-cell spread, and inflammatory destruction [28]). The zebrafish model has since furthered our understanding of the cell biology underlying *Shigella* infection, highlighting the importance of autophagy in host defence [28], as well as the role of septins in controlling inflammation [29–31].

A key advantage of the zebrafish model is its high-throughput nature, which has allowed for the study of emerging and epidemic *Shigella* infections using relevant clinical isolates. Most other model systems of shigellosis have strictly used laboratory-adapted strains (Table 1) but epidemiological studies have shown that these isolates are no longer reflective of current *Shigella* epidemiology, especially in the setting of rapidly increasing antimicrobial resistance [6,32]. A unique feature of the model is the amenability of zebrafish to incubation at different temperatures. This was most recently exploited by the zebrafish infection of a recently emerged **EIEC** clone at both 28.5°C and 32.5°C [33]. This work illuminated a role for the acquisition of pINV (and the T3SS) in shaping its thermoregulated virulence and delivered insights into the emergence of novel EIEC and *Shigella* species. In addition, the zebrafish model has been used to show that isolates of *Shigella* which are associated with long-term carriage in humans (especially within the MSM community) can establish persistent infections in zebrafish (Figure 2B) in an **O-antigen**-dependent manner [34]. Given recent advances in the field, the zebrafish model is now in a strong position to model host and bacterial factors contributing to **bacterial persistence** *in vivo* [34–36].

The zebrafish model has previously been used to investigate novel methods of *Shigella* control, including the use of the predatory bacterium ***Bdellovibrio*** [37] and, more recently, to demonstrate the efficacy of a **bacteriophage**-delivered CRISPR Cas9 system that specifically targets *S. flexneri* [38]. Additionally, high-throughput screening methods revealed that antibiotics synergise with neutrophils to control *Shigella* infection *in vivo* [39]. Another key attribute of the zebrafish larval model is the absence of an adaptive immune system for the first 4 weeks of life, allowing for the study of the innate immune system in isolation. Studies have shown that exposure to a low dose of *Shigella* can train the zebrafish innate immune system, generating protective neutrophils through epigenetic modifications, which later assist in protection against a secondary re-infection [40,41]. This model can ultimately be used to discover mechanisms underlying **trained innate immunity** [40,42] and may in the future guide *Shigella* vaccine strategies that consider both innate and adaptive immune responses.

Taken together, the zebrafish is an important tool for studying the cell biology, epidemiology, and infection control of *Shigella in vivo*. We suggest that modelling of *Shigella* infection using zebrafish

Table 1. *Shigella* strains widely used as laboratory reference strains

Species	*Shigella boydii*	*Shigella dysenteriae*		*Shigella flexneri*		*Shigella sonnei*
Strain	Sb227	Sd197	Sd1617	M90T	2457T	53G
Serotype/ lineage	Serotype 4	Serotype 1	Serotype 1	Serotype 5a	Serotype 2a	Lineage 2
Year isolated	1950s	1950s	1968	1955	1954	1954
Location isolated	China	China	Guatemala	Mexico	Japan	Japan
Genome accession	NC_007613	NC_007606	CP006736	NZ_CP037923	NC_004741	NC_016822

is not meant to replace traditional vertebrate infection models (such as mice, guinea pigs, rabbits, or primates) but can instead synergise with them to reveal fundamental concepts underlying *Shigella* pathogenesis and its control.

Inflammasome-deficient mice are susceptible to *Shigella* infection

While our knowledge of other infectious disease has flourished due to the availability of a murine model (due to an abundance of genetic tractability tools and systems), the natural resistance of mice to oral challenge with *Shigella* has significantly hampered our understanding of its pathogenesis. To overcome these barriers, the pre-treatment of mice with antibiotics, the use of alternative infection sites, and genetic modification strategies have been employed. The streptomycin-mouse model has been used to show that the *S. sonnei* **effector proteins**, OspC1 and OspC3, can block **interferon** signalling as a virulence strategy [43]. An intraperitoneal mouse model has also been used to identify the prodrug Tebipenem pivoxil as a safe and effective treatment for multidrug-resistant *Shigella* infections [44]. Despite the lack of clinical symptoms in these models, they still prove useful for studying fundamental host–pathogen interactions and identifying potential therapeutic options.

For many years, the reasons underlying the resistance of mice to *Shigella* were unknown, until recently, when the NAIP-NLRC4 **inflammasome** was implicated as a protective barrier against *Shigella* infection in mice [45]. This work revealed that *Shigella* can elude the human NAIP-NLRC4 inflammasome as an immune-evasion strategy, but in mice the T3SS and lipopolysaccharide (LPS) components of *Shigella* are recognised, leading to activation of the mouse inflammasome, resulting in pyroptosis. These findings led to the implementation of a revolutionary NAIP-NLRC4 inflammasome-deficient mouse model, which recapitulates many aspects of shigellosis upon oral challenge [45]. This model has additionally illuminated caspases-1, -8, and -11 as protective agents against *Shigella* infection, showing that their genetic removal renders mice susceptible to shigellosis [46]. Indicative of a molecular arms race between host and pathogen, it was found that *Shigella* encodes effector proteins which can antagonise this protective effect in humans. The model has subsequently been used to examine this molecular arms race, where it was used to address how *Shigella* evades pyroptosis in humans. It was revealed that the bacterial effector IpaH7.8 (a bacterial ubiquitin ligase) targets **Gasdermin D** for proteasomal degradation, ultimately blocking pyroptosis and allowing *Shigella* to replicate intracellularly, spreading from cell to cell, causing clinical symptoms of shigellosis (Figure 2C) [47]. The advances made using this model have delivered major contributions to the field of inflammasome biology [48] and towards understanding the barriers to *Shigella* infection, and in future will likely be used to study additional host and bacterial factors underlying *Shigella* pathogenesis.

An important limitation of this mouse model is the prerequisite of antibiotic treatment, which means that *Shigella* infection in the context of microbial communities cannot be fully studied. Given that *S. sonnei* is known to encode interbacterial competition weapons, such as colicins [49] and a **type six secretion system** (**T6SS**) [49,50], this is an imperative consideration because the interactions of *Shigella* with the gut microbiota are likely to shape its pathogenesis. Taken together, the timely availability of a relevant murine model can be expected to deliver many significant advances in our understanding of shigellosis and how to prevent it.

Shigella infection: the human model

Intestine-Chip, zebrafish, and inflammasome-deficient mice all represent innovative models of *Shigella* infection, each with their own unique advantages and limitations (Figure 1). Despite recent advances, none of the models can recapitulate fully the complex interaction of *Shigella* with its highly specific human host. Controlled human infection models (CHIMs), in which healthy,

consenting humans are challenged with an infectious organism in a controlled environment, are especially important for human-restricted pathogens such as *Shigella*. The immune response to *Shigella* infection has often been extrapolated from animal models, but the implementation of CHIMs offers a unique situation in which the precise human immunological response to infection can be characterised.

The *Shigella* CHIM was first implemented in 1946, where prison inmates in Illinois were infected with *S. flexneri* 2547T to evaluate the efficacy of a polyvalent *Shigella* vaccine [51]. Human challenge models of *Shigella* were then used throughout the 1970s–1980s to investigate basic *Shigella* pathology (such as the infectious dose) and test additional vaccine candidates (reviewed in [52]). The use of *Shigella* CHIMs then decreased in popularity, likely due to ethical concerns, particularly surrounding the involvement of vulnerable populations such as prison inmates. Interest in the *Shigella* CHIMs re-emerged in recent years, and in 2020 a CHIM was established using lyophilised *S. sonnei* 53G, with the aim to first standardise the *Shigella*-CHIM protocol [53] and then to characterise the human immune response to *S. sonnei* challenge [54]. These studies revealed that disease progression was linked to an intestinal inflammatory response (in agreement with results from animal models) and implicated **IgA** as a potential correlate of immunity. This represents a landmark guiding the design of future studies for establishing controlled human infection, where eventually it would be of interest to test more epidemiologically relevant isolates.

The *S. sonnei*-CHIM tested the efficacy of immunisation with 1790GAHB (a potential vaccine candidate) in a Phase 2b clinical trial, and revealed that the candidate was safe and produced **IgG** responses against *S. sonnei* LPS, but did not demonstrate significant protection against developing shigellosis [55]. Despite the shortcomings of this particular vaccine candidate, this study demonstrates the utility of a human challenge model to enable vaccine development, and allows resources to be directed towards the most promising candidates. Another CHIM, this time challenging participants with *S. flexneri* 2457T, was used to test the safety and immunogenicity of Flexyn2a, a bioconjugate vaccine candidate. This study was largely successful, verifying that the vaccine was safe and demonstrated an association between vaccine-induced IgG response and disease reduction. Since the CHIM study, Flexyn2a has been shown to elicit a long-term memory B cell response and is currently undergoing trials in both Kenya and North America [56]. Overall, this candidate, with the aid of CHIMs, represents hope for the imminent development of a vaccine against *S. flexneri*, although the heterogeneity of *Shigella* serotypes is likely to complicate this progress.

The need for a vaccine is most obvious in LMIC settings, and especially for children under 5 years of age; however, most CHIMs are performed in HIC settings, where *Shigella* infections are largely self-limiting. It has been suggested that the outcome of performing a CHIM in these regions may produce substantially different results than if it were to be performed in a more immunologically representative population where *Shigella* is endemic and causing the greatest burden of disease (reviewed in [57]). The extensive ethical considerations of employing CHIMs in the relevant populations (given that shigellosis is more severe in children, who produce a different immunological response to adults [58]), along with the complexity and cost of setting up a CHIM present as clear limitations to the model over other infection models. Considering the valuable information that can be gathered from CHIMs, the implementation of CHIMs in immunologically relevant populations needs to be carefully designed and standardised. In the future, working with CHIMs, in combination with monitoring of the changing *Shigella* epidemiological landscape [32,59,60], can help us to understand, and ultimately limit, the burden of shigellosis in the most vulnerable populations.

Concluding remarks

As the emergence of multi- and extensively-drug-resistant *Shigella* increases, the need to develop novel therapeutics and preventatives is urgent. This will require an enhanced understanding of *Shigella* infection, for which model systems are essential. Here we overview innovative technologies and animal models which, in recent years, have enabled fundamental discovery and have been used as platforms to test alternative treatments. The advances in modelling *Shigella* infection *in vitro* have revealed the importance of mimicking a mechanically realistic environment. The zebrafish model has proved versatile, and its high throughput nature has enabled the study of *Shigella* isolates currently circulating within the human population, illuminating the emergence of novel *Shigella* and EIEC species and highlighting a role for niche occupation. A revolutionary inflammasome-deficient mouse model has uncovered key barriers to *Shigella* infection, which may next be exploited to reveal determinants underlying *Shigella* pathogenesis. Finally, CHIMs have demonstrated their unique ability to provide a realistic response to a human-adapted pathogen and test the safety and efficacy of promising vaccine candidates.

Lessons learned from each these models reveal that no one model alone is sufficient to tackle the increasing burden shigellosis exerts globally (see Outstanding questions). We propose that a coordinated, interdisciplinary, and international effort using the breadth of *Shigella* infection models will be required to fully understand the complexity of *Shigella* and its interactions with its highly specific human host.

Outstanding questions

The Intestine-Chip model has advanced our fundamental understanding of the early steps of *Shigella* infection. How will the use of organoid and organ-on-a-chip models update long-standing results derived from monolayers of immortalised cells?

The use of zebrafish has enabled high-throughout testing of relevant clinical isolates of *Shigella*. Can zebrafish be used to predict emerging and epidemic strains of *Shigella*?

The inflammasome-deficient mouse closely recapitulates shigellosis for the first time in a murine model. Can protective features of mice be exploited to control *Shigella* infection in humans?

CHIMs have proved useful for testing promising vaccine candidates. Can CHIMs ethically move away from the use of laboratory adapted *Shigella* strains, considering the extensive drug resistance frequently found in contemporary isolates?

The modelling of *Shigella* infection has proved essential for understanding basic pathogenesis. How can the use of different *Shigella* models be used in synergy to combat *Shigella* infection?

Acknowledgments

We thank members of the Holt and Mostowy laboratories for helpful discussions. S.L.M. is supported by a Biotechnology and Biological Sciences Research Council LIDo PhD studentship (BB/T008709/1). Work in the S.M. laboratory is supported by a Wellcome Trust Senior Research Fellowship (206444/Z/17/Z) and European Research Council Consolidator Grant (grant agreement No. 772853-ENTRAPMENT).

Declaration of interests

No interests are declared.

References

1. Khalil, I.A. *et al.* (2018) Morbidity and mortality due to *Shigella* and enterotoxigenic *Escherichia coli* diarrhoea: the Global Burden of Disease Study 1990–2016. *Lancet Infect. Dis.* 18, 1229–1240
2. Kotloff, K.L. *et al.* (2018) Shigellosis. *Lancet* 391, 801–812
3. Kotloff, K.L. *et al.* (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. *Lancet* 382, 209–222
4. Kotloff, K.L. *et al.* (2012) The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study. *Clin. Infect. Dis.* 55, S232–S245
5. Baker, K.S. *et al.* (2015) Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: a cross-sectional study. *Lancet Infect. Dis.* 15, 913–921
6. Mason, L.C.E. *et al.* (2023) The evolution and international spread of extensively drug resistant Shigella sonnei. *Nat. Commun.* 14, 1983
7. Bardsley, M. *et al.* (2020) Persistent transmission of shigellosis in England is associated with a recently emerged multidrug-resistant strain of *Shigella sonnei*. *J. Clin. Microbiol.* 58, e01692–19
8. Alphonse, N. and Odendall, C. (2023) Animal models of shigellosis: a historical overview. *Curr. Opin. Immunol.* 85, 102399
9. Sereny, B. (1955) Experimental shigella keratoconjunctivitis; a preliminary report. *Acta Microbiol. Acad. Sci. Hung.* 2, 293–296
10. Sansonetti, P.J. *et al.* (1982) Involvement of a plasmid in the invasive ability of *Shigella flexneri*. *Infect. Immun.* 35, 852–860
11. Watanabe, H. and Nakamura, A. (1985) Large plasmids associated with virulence in *Shigella* species have a common function necessary for epithelial cell penetration. *Infect. Immun.* 48, 260–262
12. Martyn, J.E. *et al.* (2022) maintenance of the *Shigella sonnei* virulence plasmid is dependent on its repertoire and amino acid sequence of toxin–antitoxin systems. *J. Bacteriol.* 204, e0051921
13. Shim, D.-H. *et al.* (2007) New animal model of shigellosis in the Guinea pig: its usefulness for protective efficacy studies. *J. Immunol.* 178, 2476–2482
14. Perdomo, O.J. *et al.* (1994) Acute inflammation causes epithelial invasion and mucosal destruction in experimental shigellosis. *J. Exp. Med.* 180, 1307–1319
15. Sansonetti, P.J. *et al.* (1996) Infection of rabbit Peyer's patches by *Shigella flexneri*: effect of adhesive or invasive bacterial phenotypes on follicle-associated epithelium. *Infect. Immun.* 64, 2752–2764
16. Zychlinsky, A. *et al.* (1992) *Shigella flexneri* induces apoptosis in infected macrophages. *Nature* 358, 167–169
17. Fernandez-Prada, C.M. *et al.* (2000) *Shigella flexneri* IpaH(7.8) facilitates escape of virulent bacteria from the endocytic vacuoles of mouse and human macrophages. *Infect. Immun.* 68, 3608–3619
18. Schnupf, P. and Sansonetti Philippe, J. (2019) *Shigella* pathogenesis: new insights through advanced methodologies. *Microbiol. Spectr.* 7 (2). https://doi.org/10.1128/microbiolspec.bai-0023-2019
19. Bernardini, M.L. *et al.* (1989) Identification of icsA, a plasmid locus of *Shigella flexneri* that governs bacterial intra- and intercellular spread through interaction with F-actin. *Proc. Natl. Acad. Sci. USA* 86, 3867–3871

20. Egile, C. *et al.* (1999) Activation of the CDC42 effector N-WASP by the *Shigella flexneri* IcsA protein promotes actin nucleation by Arp2/3 complex and bacterial actin-based motility. *J. Cell Biol.* 146, 1319–1332
21. Mostowy, S. *et al.* (2010) Entrapment of intracytosolic bacteria by septin cage-like structures. *Cell Host Microbe* 8, 433–444
22. Krokowski, S. *et al.* (2018) Septins recognize and entrap dividing bacterial cells for delivery to lysosomes. *Cell Host Microbe* 24, 866–874.e4
23. Ranganathan, S. *et al.* (2019) Evaluating *Shigella flexneri* pathogenesis in the human enteroid model. *Infect. Immun.* 87, e00740-18
24. Perlman, M. *et al.* (2023) A foundational approach to culture and analyze malnourished organoids. *Gut Microbes* 15, 2248713
25. Grassart, A. *et al.* (2019) Bioengineered human organ-on-chip reveals intestinal microenvironment and mechanical forces impacting *Shigella* infection. *Cell Host Microbe* 26, 435–444.e4
26. Boquet-Pujadas, A. *et al.* (2022) 4D live imaging and computational modeling of a functional gut-on-a-chip evaluate how peristalsis facilitates enteric pathogen invasion. *Sci. Adv.* 8, eabo5767
27. Prouty, M.G. *et al.* (2003) *Zebrafish–Mycobacterium marinum* model for mycobacterial pathogenesis. *FEMS Microbiol. Lett.* 225, 177–182
28. Mostowy, S. *et al.* (2013) The zebrafish as a new model for the *in vivo* study of *Shigella flexneri* interaction with phagocytes and bacterial autophagy. *PLoS Pathog.* 9, e1003588
29. Mazon-Moya, M.J. *et al.* (2017) Septins restrict inflammation and protect zebrafish larvae from *Shigella* infection. *PLoS Pathog.* 13, e1006467
30. Torraca, V. *et al.* (2023) Zebrafish null mutants of Sept6 and Sept15 are viable but more susceptible to *Shigella* infection. *Cytoskeleton (Hoboken)* 80, 266–274
31. Van Ngo, H. *et al.* (2023) Septins promote caspase activity and coordinate mitochondrial apoptosis. *Cytoskeleton* 80, 254–265
32. Chung The, H. *et al.* (2021) Evolutionary histories and antimicrobial resistance in *Shigella flexneri* and *Shigella sonnei* in Southeast Asia. *Commun. Biol.* 4, 353
33. Miles, S.L. *et al.* (2023) Acquisition of a large virulence plasmid (pINV) promoted temperature-dependent virulence and global dispersal of O96:H19 enteroinvasive *Escherichia coli*. *mBio* 14, e00882-23
34. Torraca, V. *et al.* (2023) *Shigella* serotypes associated with carriage in humans establish persistent infection in zebrafish. *J. Infect. Dis.* 228, 1108–1118
35. Jade, L. *et al.* (2024) Dynamics of macrophage polarization support *Salmonella* persistence in a whole living organism. *Elife* 13, e89828
36. Stuti, K.D. *et al.* (2023) RpoS activates *Salmonella* Typhi biofilms and drives persistence in a Zebrafish model. *bioRxiv*, Published online October 26, 2023. https://doi.org/10.1101/2023.10.26.564249
37. Willis, A.R. *et al.* (2016) Injections of predatory bacteria work alongside host immune cells to treat *Shigella* infection in zebrafish larvae. *Curr. Biol.* 26, 3343–3351
38. Huan, Y.W. *et al.* (2023) P1 bacteriophage-enabled delivery of CRISPR-Cas9 antimicrobial activity against *Shigella flexneri*. *ACS Synth. Biol.* 12, 709–721
39. Lensen, A. *et al.* (2023) An automated microscopy workflow to study *Shigella*–neutrophil interactions and antibiotic efficacy *in vivo*. *Dis. Model. Mech.* 16, dmm049908
40. Gomes, M.C. *et al.* (2023) *Shigella* induces epigenetic reprogramming of zebrafish neutrophils. *Sci. Adv.* 9, eadf9706
41. Willis, A.R. *et al.* (2018) *Shigella*-induced emergency granulopoiesis protects zebrafish larvae from secondary infection. *mBio* 9, e00933–18
42. Darroch, H. *et al.* (2023) Infection-experienced HSPCs protect against infections by generating neutrophils with enhanced mitochondrial bactericidal activity. *Sci. Adv.* 9, eadf9904
43. Alphonse, N. *et al.* (2022) A family of conserved bacterial virulence factors dampens interferon responses by blocking calcium signaling. *Cell* 185, 2354–2369.e17
44. Fernández Álvaro, E. *et al.* (2022) The repurposing of Tebipenem pivoxil as alternative therapy for severe gastrointestinal infections caused by extensively drug-resistant *Shigella* spp. *Elife* 11, e69798
45. Mitchell, P.S. *et al.* (2020) NAIP–NLRC4-deficient mice are susceptible to shigellosis. *eLife* 9, e59022
46. Roncaioli, J.L. *et al.* (2023) A hierarchy of cell death pathways confers layered resistance to shigellosis in mice. *eLife* 12, e83639
47. Luchetti, G. *et al.* (2021) *Shigella* ubiquitin ligase IpaH7.8 targets gasdermin D for degradation to prevent pyroptosis and enable infection. *Cell Host Microbe* 29, 1521–1530.e10
48. Rojas-Lopez, M. *et al.* (2023) NLRP11 is a pattern recognition receptor for bacterial lipopolysaccharide in the cytosol of human macrophages. *Sci. Immunol.* 8, eabo4767
49. De Silva, P.M. *et al.* (2023) *Escherichia coli* killing by epidemiologically successful sublineages of *Shigella sonnei* is mediated by colicins. *EBioMedicine* 97, 104822
50. Anderson, M.C. *et al.* (2017) *Shigella sonnei* encodes a functional *T6SS* used for interbacterial competition and niche occupancy. *Cell Host Microbe* 21, 769–776.e3
51. Shaughnessy, H.J. *et al.* (1946) Experimental human bacillary dysentery: polyvalent dysentery vaccine in its prevention. *JAMA J. Am. Med. Assoc.* 132, 362–368
52. Porter, C.K. *et al.* (2013) The *Shigella* human challenge model. *Epidemiol. Infect.* 141, 223–232
53. Frenck Robert, W. *et al.* (2020) Establishment of a controlled human infection model with a lyophilized strain of *Shigella sonnei* 53G. *mSphere* 5, e00416–20
54. Clarkson, K.A. *et al.* (2020) Immune response characterization after controlled infection with lyophilized *Shigella sonnei 53G*. *mSphere* 5, e00988–19
55. Frenck, R.W. *et al.* (2021) Efficacy, safety, and immunogenicity of the *Shigella sonnei* 1790GAHB GMMA candidate vaccine: results from a phase 2b randomized, placebo-controlled challenge study in adults. *EClinicalMedicine* 39, 101076
56. Meron-Sudai, S. *et al.* (2023) A *Shigella flexneri* 2a synthetic glycan-based vaccine induces a long-lasting immune response in adults. *NPJ Vaccines* 8, 35
57. Giersing, B.K. *et al.* (2019) How can controlled human infection models accelerate clinical development and policy pathways for vaccines against *Shigella*? *Vaccine* 37, 4778–4783
58. Raqib, R. *et al.* (2000) Innate immune responses in children and adults with shigellosis. *Infect. Immun.* 68, 3620–3629
59. Vinh, H. *et al.* (2009) A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation. *BMC Infect. Dis.* 9, 204
60. Thompson, C.N. *et al.* (2015) The rising dominance of *Shigella sonnei*: an intercontinental shift in the etiology of bacillary dysentery. *PLoS Negl. Trop. Dis.* 9, e0003708

1.6. Zebrafish as a model for studying host-pathogen interactions

Mice are naturally resistant to oral challenge with *Shigella* in part due to differences in inflammasome biology (Mitchell et al., 2020), a fact that has hampered the study of *Shigella* in an *in vivo* context. Zebrafish (*Danio rerio*) have emerged as a valuable model to study *Shigella* infection, delivering important insights into bacterial drivers of virulence (Torraca et al., 2019, Virgo et al., 2024, Torraca et al., 2023), and the host response to infection (Willis et al., 2018, Lensen et al., 2023, Gomes et al., 2023, Mazon-Moya et al., 2017). This sub-section will consider the advantages and suitability of using zebrafish as a tool for modelling infection in greater detail. In contrast to many other infection models, zebrafish are highly fecund and breed quickly with a typical clutch consisting of more than 100 embryos (Goldsmith and Jobin, 2012). In addition to this, embryos develop rapidly (Fig 1.5) and reach sexual maturity in around 3 months, providing a relatively fast, reliable and high-throughput system for studying infection biology.

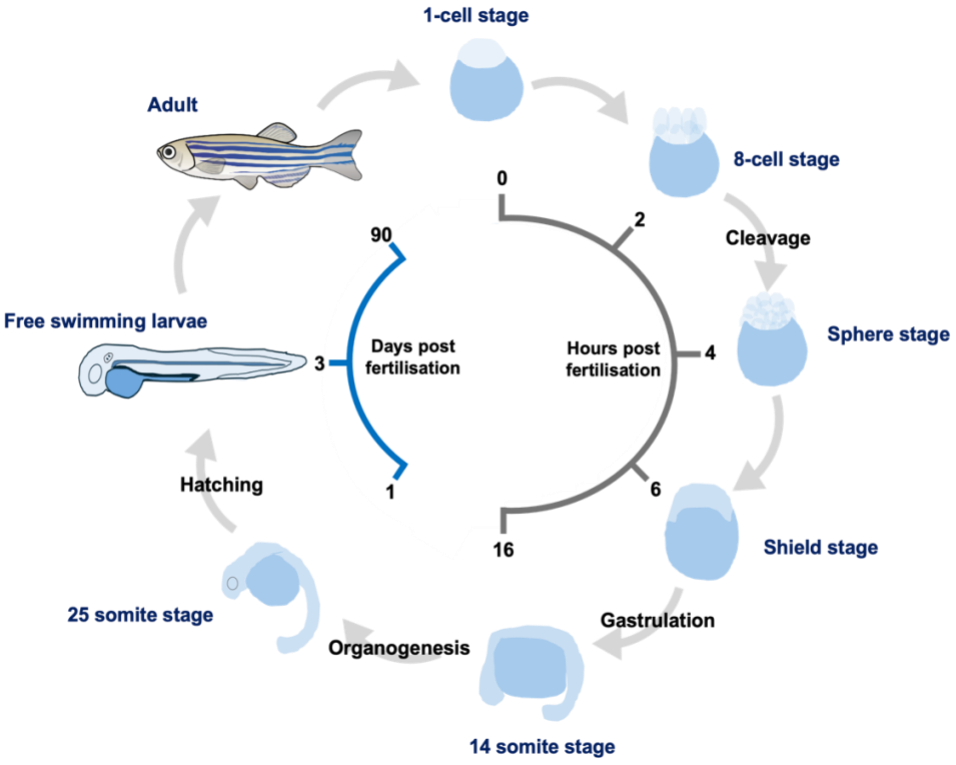


Figure 1.5. Life cycle of the zebrafish. Zebrafish undergo rapid development following fertilisation, development starts at the 1-cell stage and within 48-72 hours, the larvae are free swimming. A functional innate immune system is present by around 33 hours post fertilisation, where neutrophils and macrophages are phagocytic and capable of destroying pathogens. Zebrafish are considered adults by around 90 days after fertilisation, where they also reach sexual maturity. Figure adapted from D'Costa and Shepherd, 2009.

Sequencing of the zebrafish genome revealed that around 70% of genes are orthologous to those present in humans, and this rises to 82% when only considering genes that are implicated in human disease (Howe et al., 2013). In line with this, the zebrafish immune system is similar to the mammalian immune system, comprised of both an innate and adaptive arm in fully developed fish (Renshaw and Trede, 2012). In larval and juvenile fish (up to ~30 days post fertilisation (dpf)), only the innate arm is functional, enabling the study of response to infection without the influence of the adaptive arm. Immune cells in the zebrafish larvae are homologous, in terms of both phenotype and function, to those in humans. The neutrophil is the most abundant innate immune cell in zebrafish and like in humans, can be identified

through the expression of *myeloid-specific peroxidase* (*mpx*) or *lysozyme C (lyz)* (Renshaw et al., 2006). Similarly, macrophages can be distinguished through the expression of *macrophage-expressed gene 1 (mpeg)* (Ellett et al., 2011). A further advantage of using zebrafish larvae is their transparency in the first weeks of life and genetically tractability, which enables the coupling of immune cells to fluorescent reporter proteins for visualisation.

Immune cell myelopoiesis can be detected as early as 12 hours post fertilisation (hpf) with macrophages differentiating first, followed by neutrophil differentiation by 33 hpf (Le Guyader et al., 2008). Importantly, zebrafish neutrophils and macrophages are fully functional in the early stages of development, capable of phagocytosis, the generation of respiratory bursts and performing NETosis (the formation of neutrophil extracellular traps which enables the killing of microbes) (Palić et al., 2007). Overall, the zebrafish model offers an opportunity to study host-pathogen interactions at the whole animal level and resolve bacterial and host drivers of virulence in fine detail.

1.7. Summary and aims

Shigella and EIEC represent a diverse pathotype of *E. coli* that have evolved to become human adapted pathogens, responsible for a significant proportion of diarrhoeal morbidity and mortality globally (Kotloff et al., 2012). Through the process of convergent evolution, several acquisitions (including a virulence plasmid encoding a T3SS among other factors) and losses (including antivirulence genes, metabolic pathways and immunogenic factors) have been documented in the stepwise adaptation to the human host. The zebrafish larva is a versatile model, useful for studying host-pathogen interactions and uncovering drivers of bacterial virulence. In this thesis, I use a combination of zebrafish infection and bacterial genomics to understand both functional and genomic signatures of epidemiological success at various stages of *Shigella* evolution. Specific aims can be found below.

1. Use the *Shigella*-zebrafish model to investigate the early stages of *Shigella* and EIEC emergence exploiting a novel clone of EIEC (**Chapter 3**).

2. Generate and compare reference genomes for epidemiologically important variants of *S. sonnei,* the highest burden *Shigella* species in HIC and economically transitioning countries (**Chapter 4**).

3. Use the *Shigella*-zebrafish infection model to explore the role of virulence in pathogen success and uncover signatures of epidemiological success within *S. sonnei* (**Chapter 5**).

Chapter 2. Materials and methods

2.1. Bacterial methodologies

2.1.1. Bacterial strains and growth conditions

Clinical isolates of EIEC and *S. sonnei* (Table 1.1) were obtained in collaboration with the United Kingdom Health and Security Agency (UKHSA), Institut Pasteur, The University of Liverpool (UoL) and Oxford University Clinical Research Unit (OUCRU) and were originally collected as part of routine public health surveillance. Strains were received on agar slants and streaked on plates of Tryptic Soy Agar (TSA; Sigma Aldrich), supplemented with 0.01% Congo Red (CR; Sigma Aldrich) to select for pINV+ isolates. Red colonies, which indicate T3SS and pINV presence, were picked, stored in glycerol at 25% (v/v) and stored at -80 °C. Overnight cultures were prepared by inoculating 5 mL Trypticase Soy Broth (TSB; Sigma Aldrich), where necessary supplemented with 100 µg/mL carbenicillin, with a single red colony and incubating at 37 °C for ~16 hours, with shaking at 200 rotations per minute (rpm).

Table 2.1. Details of bacterial strains used in this study. For ease of understanding, *S. sonnei* strains are referred to by genotype instead of strain ID throughout. Carb = carbenicillin, Kan = kanamycin, ^{R}= resistant. UOL = University of Liverpool, OUCRU = Oxford University Clinical Research Unit.

Species	Strain ID	Description	Serotype/genotype	Obtained from	Source
E. coli	NCTC 9096	pINV- ST99	O96:H19	NCTC	(Miles et al., 2023)
	pINV+1	pINV+ ST99 EIEC, clinical isolate	O96:H19	UKHSA	(Miles et al., 2023)

	pINV+2	pINV+ ST99 EIEC, clinical isolate	O96:H19	UKHSA	(Miles et al., 2023)
	pINV+3	pINV+ ST99 EIEC, clinical isolate	O96:H19	UKHSA	(Miles et al., 2023)
	pINV+4	pINV+ ST99 EIEC, clinical isolate	O96:H19	UKHSA	(Miles et al., 2023)
S. flexneri	M90T	Lab strain	5a	Mostowy Lab	(Mostowy et al., 2010)
S. sonnei	53G	Lab strain	2.8	Mostowy Lab	(Torraca et al., 2019)
	53G Δ*waaI*	O antigen mutant, $Carb^R$, Kan^R	2.8	Mostowy Lab	(Torraca et al., 2019)
	53G Δ*g4c*	Capsule mutant, $Carb^R$	2.8	Mostowy Lab	(Torraca et al., 2019)
	53G GFP	Transformed with pFPV25.1, GFP reporter, $Carb^R$	2.8	Mostowy Lab	(Torraca et al., 2019)
	02-1157 GFP	Transformed with pFPV25.1, GFP reporter, $Carb^R$	3.6.1.1.1	This study	NA
	03-0142 GFP	Transformed with pFPV25.1, GFP reporter, $Carb^R$	3.7.29.1	This study	NA
	201809330	Clinical isolate	1	Institut Pasteur	NA
	SRR15429165	Clinical isolate	1.5	UKHSA	NA

	SRR7209033	Clinical isolate	2.1	UKHSA	NA
	SRR10380789	Clinical isolate	2.3	UKHSA	NA
	SRR7286515	Clinical isolate	2.12.4	UKHSA	NA
	SRR7826923	Clinical isolate	3.4.1	UOL	NA
	SRR8086736	Clinical isolate	3.6.1	UKHSA	NA
	02-1157	Clinical isolate	3.6.1.1.1	OUCRU	NA
	SRR8240966	Clinical isolate	3.6.2	UOL	NA
	SRR8166038	Clinical isolate	3.7.11	UKHSA	NA
	SRR7866101	Clinical isolate	3.7.16	UKHSA	NA
	SRR8032849	Clinical isolate	3.7.28	UKHSA	NA
	03-0142	Clinical isolate	3.7.29.1.2	OUCRU	NA
	SRR8114786	Clinical isolate	3.7.30.1	UOL	NA
	SRR7367430	Clinical isolate	3.7.30.4.1	UOL	NA
	SRR8426575	Clinical isolate	3.7.25	UOL	NA
	SRR7291665	Clinical isolate	3.7.3	UKHSA	NA
	SRR5005344	Clinical isolate	3.7.30.4	UKHSA	NA

2.1.2. Generating fluorescent bacteria

Fluorescent bacterial strains were generated by transformation with pFPV25.1 plasmid, which encodes a green fluorescent protein (GFP) reporter and resistance to carbenicillin as a selective marker. Electrocompetent cells were generated by growing bacteria to an optical density (OD) of 0.3-0.4, centrifugation at 4 °C and washing with 10% (v/v) ice cold glycerol. 100 ng of DNA was added, and the suspension was electroporated using the Ec2 setting on a Bio-Rad Micropulser. 1 mL of TSB was immediately added, and bacteria were left to recover

at 37 °C for 2 hours. Following recovery, the suspension was plated on TSA supplemented with 50 µg/mL carbenicillin to select for positive colonies.

2.1.3. Growth curves

100 µL of overnight culture was diluted in 10 mL TSB and 100 µL of culture in 900 µL of TSB was used to generate OD measurements (at absorbance 600 nm), which were taken every 30 minutes. For ODs higher than 1, the sample was diluted and remeasured to get an accurate reading. For acid tolerance experiments, the pH was adjusted to pH 5 using a few drops of concentrated hydrochloric acid (Sigma Aldrich).

2.1.4. Serum survival assay

One fresh red colony was inoculated into 50 µL PBS. Lyophilised baby rabbit complement (Bio-Rad, C12CA) was resuspended in 2 mL ice cold water and 150 µL was added to the bacterial suspension. 10 µL was taken for serial dilution and plating to determine the initial inoculum. The bacterial-rabbit complement mixture was incubated at 37 °C, shaking at 200 rpm for 4 hours before another 10 µL was taken, serially diluted and plated. CR+ colonies were counted the following morning and fold change CFU was calculated by normalising colonies counted at 4 hours to the initial inoculum. As a control, baby rabbit complement was heat killed at 56 °C for 30 minutes to inactivate the complement system.

2.1.5. RNA extraction and qRT-PCR

Bacterial cultures were grown to mid-exponential phase, as described above. An amount of culture corresponding to $1x10^9$ bacterial cells was pelleted by centrifugation at 5000 rpm at 4 °C for 10 minutes, the supernatant was removed, and the pellet was kept at -80 °C overnight to aid with cell lysis. RNA was then extracted using the Monarch Total RNA Miniprep Kit (New England Biolabs) as per the manufacturer's instructions. RNA concentration was measured using a DeNovix DS-11 spectrophotomer.1000 ng of RNA was then converted to cDNA using a QuantiTect reverse transcription kit (Qiagen). Template cDNA was subjected to quantitative reverse transcription PCR (qRT-PCR) using a 7500 Fast Real-Time PCR System machine and SYBR green master mix (Applied Biosystems), with samples run in technical duplicates.

Primers used can be found in Table 2.2, *rrsA* was used as a housekeeping gene (as validated previously (Guyet et al., 2018)) and the delta-delta Ct method was used to quantify gene expression.

Table 2.2. Primers used to measure bacterial gene expression via qRT-PCR.

Primer name	Sequence (5'-3')	Source
rrsA_FW	AACGTCAATGAGCAAAGGTATTAA	(Koppolu et al., 2013)
rrsA_RV	GAACTTCAAGATCTGCTCCTGC	(Koppolu et al., 2013)
etp_FW	CTCAATCCTGGTGGTTTGTACCG	(Caboni, 2013)
etp_RV	GACTCCATTGCCAGAATCAGATC	(Caboni, 2013)
etk_FW	CAGGCAGCACTCAGGAAAATGAG	(Caboni, 2013)
etk_RV	GATTGCAGCAGTTGGATCTCCG	(Caboni, 2013)
virF_FW	AAAGGTGTTCAATGACGGTTAGC	(Skovajsová et al., 2022)
virF_RV	CAATTTGCCCTTCATCGATAGTC	(Skovajsová et al., 2022)
virB_FW	GGAAGGCCAAAAGAAAGAGTTTACA	(Skovajsová et al., 2022)
virB_RV	GAGGAATCTTGGCTTTGATAAAGG	(Skovajsová et al., 2022)
ipaB_Fw	CTGCATTTTCAAACACAGC	(Koppolu et al., 2013)
ipaB_Rv	GAGTAACACTGGCAAGTC	(Koppolu et al., 2013)
wzzB_Fw	GCGATAACATTCAGGCGCAA	This study
wzzB_RV	CCCCTGGTAATGCACCAAGA	This study

2.1.6. DNA sequencing

The DNA extraction, long-read (Oxford Nanopore Technology) and short read (Illumina) sequencing of *S. sonnei* clinical isolates was performed by MicrobesNG as a paid service. To prepare isolates for sequencing, an overnight culture was grown and sub-cultured as previously described in section 2.1.1. Once the sub-culture reached mid-exponential phase,

cells were pelleted by centrifugation at 4000 rpm for 4 minutes, washed in 1 mL PBS and resuspended at a density of $4x10^9$ cells in 500 µL DNA shield (Zymo Research). For Illumina sequencing, paired-end libraries were prepared with the Nextera XT Library Prep Kit and sequenced on an Illumina NovaSeq 6000 instrument to generate 250 bp reads. For ONT sequencing, libraries were prepared with the SQK-RBK114.96 kit (Oxford Nanopore Technologies) and loaded onto an R.10.4.1 flow cell for sequencing on a GridION instrument.

2.2. Biochemical methodologies

2.2.1. Protein precipitation

To analyse the secretion of T3SS effector proteins, bacterial cultures were grown until OD 0.2-0.3 and CR was added at a concentration of 200 µg/mL. Cultures were then grown to maximum OD (2-3). Bacterial cells were pelleted by centrifugation at 10,000 x *g*, at 4 °C for 10 minutes. The supernatant was then removed and filtered using 0.2 µm filters to remove cell debris. A volume of filtered supernatant corresponding to OD 2 was taken and trichloroacetic acid (Sigma Aldrich) was added at 10% (v/v) concentration, this mixture was then incubated at -20 °C overnight. The following day, samples were pelleted at 4 °C at maximum speed and washed with 1 mL ice cold acetone (Sigma Aldrich) 3 times to remove any residual acid. Samples were resuspended in 25 µL 1X Laemmli buffer (10 mM Tris-HCl pH 6.8, 2% sodium dodecyl 63 sulphate [SDS], 10% glycerol, 5% β-mercaptoethanol, 0.01% bromophenol blue) prior to SDS-PAGE.

2.2.2. SDS-PAGE and Coomassie staining

Protein samples were boiled at 100 °C, 17.5 µL of sample was loaded onto 12% SDS polyacrylamide gels and gels were run at 100 V in 1X Tris-glycine-SDS running buffer in a Bio-Rad Protean Tetra cell. For secretion experiments, gels were rinsed with distilled water, then stained using Coomassie Brilliant Blue R-250 (Bio-Rad) overnight, before de-staining in a solution of methanol, acetic acid, and water (30%, 5% and 65% (v/v) respectively).

2.2.3. Crude lipopolysaccharide (LPS) extraction and visualisation

Crude LPS extraction was carried out as described previously (Davis and Goldberg, 2012). Bacteria were grown overnight and then sub-cultured to reach mid-exponential phase as described above. Samples were normalised by optical density to ensure an equal density of bacteria in each sample before pelleting at 10, 000 x *g* for 10 minutes at 4 °C. Pelleted bacteria were resuspended in 200 µL 1X Laemmli buffer and boiled for 15 minutes. 5 µL DNAse I and RNAse (10 mg/mL) was added, and the mixture was incubated at 37 °C for 30 minutes. 10 µL of Proteinase K (10 mg/mL) was then added and the mixture was incubated at 59 °C for a

further 3 hours. Following this, 200 µL of ice-cold Tris saturated phenol was added to each sample. Samples were then heated to 65 °C for 15 minutes, with occasional vortexing and once cool, 1 mL petroleum ether was added. Samples were centrifuged at 14,000 x *g* for 10 minutes; the aqueous layer was isolated and added to 150 µL Laemmli buffer. 10 µL was added to a 12% SDS polyacrylamide gel and run as described above.

LPS was then visualised using a modified silver stain which oxidises LPS, allowing for better visualisation (Tsai and Frasch, 1982). Briefly, following SDS-PAGE gels were rinsed with water and then incubated with a fixing solution (40% ethanol, 5% acetic acid) overnight. Fixing solution was then replaced with an oxidising solution (0.7% periodic acid, 40% ethanol, 5% acetic acid) for 20 minutes. Following the oxidation step, the gel was washed three times in distilled water for 10 minutes and was then stained using the Pierce Silver Staining Kit according to the manufacturer's directions. All gels were visualised using a ChemiDoc Touch Gel Imaging System.

2.3. Zebrafish methodologies

2.3.1. Ethics statements

Animal experiments were performed according to the Animals (Scientific Procedures) Act 1986 and approved by the Home Office (Project licenses: PPL P84A89400 and P4E664E3C). All zebrafish experiments were performed on larvae up to 5 days post fertilisation (dpf).

2.3.2. Zebrafish husbandry

Zebrafish embryos were obtained from naturally spawning larvae and incubated at 28.5 °C in 0.5 x E2 medium (15 mM NaCl, 1 mM MgSO4, 500 µM KCl, 150 µM KH2PO4, 50 µM Na2HPO4, 0.3 µg/ml methylene blue). Zebrafish strains used for breeding are indicated in Table 2.3.

Table 2.3. Details of zebrafish lines used in this thesis.

Line	Reporter	Source
Wild type AB	N/A	NA
Tg(*mpx*::GFP)^{i144}	Neutrophils	(Renshaw et al., 2006)
Tg(*mpeg1:Gal4-FF*)^{gl25}/Tg(UAS:LIFEACT-GFP)^{mu271}	Macrophages	(Ellett et al., 2011)

2.3.3. Injection inoculum preparation

20 mL of TSB was inoculated with 400 µL overnight culture and grown to mid-exponential phase. Bacteria was harvested by centrifugation (4000 x *g*, 5 minutes), washed in 1 mL phosphate-buffered saline (PBS; Sigma Aldrich) to remove residual media and pelleted again (1 minute, 6000 x *g*). The desired inoculum concentration was achieved by measuring the OD of bacteria and correction to the desired OD. Injection inoculum was prepared by resuspension of bacteria in inoculum buffer (2% polyvinyl-pyrrolidone (PVP; Sigma Aldrich), PBS and 0.5%

phenol red (Sigma Aldrich)) to a final volume of 100 µL. Control groups of larvae were injected with just inoculum buffer.

2.3.4. Zebrafish infection

Larvae were anesthetised with tricaine (200 µg/mL; Sigma Aldrich) and placed on a 1% agarose (Bioline) pad for injections. Injections were performed in the hindbrain ventricle site (Fig 2.1) at 3 days post fertilisation (dpf). Injections were performed using borosilicate glass capillaries prepared with a P-97 Flaming / Brown Micropipette Puller (Sutter Instruments) and opened manually. 1 nL of inoculum or inoculum buffer was injected into a single larva using an IM-300 microinjector (Narishige) and an M-152 micromanipulator. Following injection, larvae were placed in E2 medium to recover from anaesthesia.

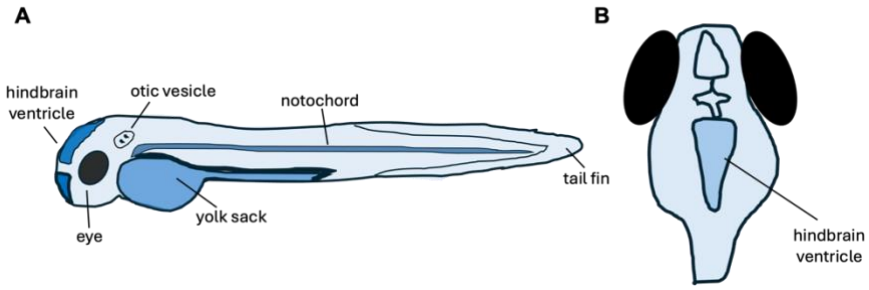


Figure 2.1. Schematic of zebrafish larvae anatomy. A) A schematic of a zebrafish larvae on its lateral side. **B)** A schematic of the zebrafish head depicting the hindbrain ventricle infection site.

2.3.5. Quantification of inoculum and bacterial burden

The precise inoculum was determined retrospectively by the mechanical disruption of a single larva in 200 µL 0.4% Triton-X-100 (Sigma Aldrich) at 0 hours post infection (hpi), and at 6 and 24 hpi for bacterial burden quantifications. For each time point, four different larvae were selected at random as representatives for the infected population. Larvae homogenates were serially diluted in PBS, plated on TSA plates supplemented with CR and colonies were counted manually following overnight incubation at 37 °C.

2.3.6. Survival assays

Larvae were maintained in groups of two or three in 24-well plates at 28.5 °C or 32.5 °C and visualised using a Leica KL300 LED microscope to check survival. Survival was assessed at

24 hpi and 48 hpi and determined via the presence or absence of a heartbeat within a 30 second period.

2.3.7. RNA extraction and qRT-PCR

For each condition, ~15 embryos were pooled and frozen overnight at -80 °C. RNA was then extracted using the RNAeasy Minikit (Qiagen), converted to cDNA and subjected to qRT-PCR as previously described (in section 2.1.5). Primers used for zebrafish qRT-PCR can be found in Table 2.4. Zebrafish gene *eef1a1a* was used as a housekeeping gene and the delta-delta Ct method was used to quantify changes in gene expression.

Table 2.4. Primers used to measure zebrafish gene expression.

Primer name	Sequence (5'-3')	Source
***eef1a1a*FW**	AAGCTTGAAGACAACCCCAAGAGC	(Herbst et al., 2015)
eef1a1aRV	ACTCCTTTAATCACTCCCACCGCA	(Herbst et al., 2015)
***cxcl8a*FW**	TGTGTTATTGTTTTCCTGGCATTTC	(Stockhammer et al., 2009)
***cxcl8a*RV**	GCGACAGCGTGGATCTACAG	(Stockhammer et al., 2009)
***cxcl18b*FW**	TCTTCTGCTGCTGCTTGCGGT	(Torraca et al., 2017)
***cxcl18b*RV**	GGTGTCCCTGCGAGCACGAT	(Torraca et al., 2017)
***il1b*FW**	GAACAGAATGAAGCACATCAAACC	(Stockhammer et al., 2009)
***il1b*RV**	ACGGCACTGAATCCACCAC	(Stockhammer et al., 2009)
***il10*FW**	CATAACATAAACAGTCCCTATG	(Boucontet et al., 2018)
***il10*RV**	GTACCTCTTGCATTTCACCA	(Boucontet et al., 2018)
***tnfa*FW**	AGACCTTAGACTGGAGAGATGAC	(Stockhammer et al., 2009)

***tnfa*RV**	CAAAGACACCTGGCTGTAGAC	(Stockhammer et al., 2009)

2.3.8. Dexamethasone treatment

Dexamethasone (Sigma Aldrich) was resuspended in dimethyl sulfoxide (DMSO, Sigma Aldrich) at a concentration of 25 mg/mL. Injections were performed as previously described (in section 2.3.4) and recovered larvae were split into two groups, treated and control. The treated group were incubated in E2 medium containing 50 µg/mL dexamethasone, and the control group were incubated in E2 medium with the same concentration of DMSO added.

2.3.9. Microscopy

For dissemination experiments, larvae were imaged using a Zeiss CellDiscoverer 7. For imaging, larvae were anaesthetised in 1X tricaine (Sigma Aldrich), placed into a 96-well plate (Perkin Elmer) and embedded in 1 drop of 1% low-melting point agarose (w/v, Thermo Scientific). Wells were topped up with tricaine for the duration of the imaging process.

For leukocyte counts, larvae were imaged using a Leica M205FA microscope and 10× (NA 0.5) dry objective. A 1% agarose (w/v) pad with ridges was used and larvae were mounted in the ridges manually, covered in 1% low-melting point agarose and 1X tricaine to immobilise for imaging. A heating pad was used throughout, to achieve a temperature of ~32.5 °C. For whole embryo imaging, larvae were placed on their lateral side, and for hindbrain ventricle imaging, larvae were positioned yolk-sack down using a micro loader pipette tip. Zebrafish leukocyte counts were then performed manually and image files were processed in ImageJ (v.1.54).

2.4. Eukaryotic cell methodologies

2.4.1. Cell culture conditions

HeLa cells (Human ATCC CCL-2) were used for eukaryotic cell infections. Cells were cultured in Dulbecco's Modified Eagle Medium (DMEM, Sigma Aldrich) supplemented with 10% foetal-bovine serum (FBS, Thermo Fisher Scientific) and incubated in an incubator supplied with 5% CO_2 at 37 °C.

2.4.2. HeLa cell infection

Cells were seeded at a density of $1.5x10^5$ in a 6-well plate (VWR) 48 hours before infection. Bacterial cultures were grown to mid-exponential phase and then diluted in DMEM (not supplemented with FBS) to reach a multiplicity of infection (MOI) of 100:1. 1 mL of bacterial suspension was added to each well and cells were centrifuged at 500 x *g* for 10 minutes at room temperature. Cells were then incubated for 30 minutes with 5% CO_2 at 37 °C. Bacterial suspension was removed from the cells and the cells were washed three times with PBS. Washed cells were treated with 2000 µL of 50 mg/mL gentamycin (Sigma Aldrich) in DMEM +FBS for 1 or 3 hours for invasion and replication timepoints respectively. Following treatment, cells were lysed with 0.1% Triton-X-100 (Sigma Aldrich) in PBS for 5 minutes at 37 °C. Cell lysates were then serially diluted and plated on TSA, plates were incubated overnight and CFU counts were determined the following morning.

2.4.3. Immunofluorescence assay

Cells previously seeded and infected on coverslips were washed 3 times with PBS and then fixed using 4% (v/v) paraformaldehyde (PFA) in PBS for 15 minutes at room temperature. Fixed cells were washed 3 times in PBS. Cells were stained with Hoechst 33342 and AlexaFluor-647 Phalloidin (A22287) suspended in PBS supplemented with 0.1% Triton X-100 and 1% bovine serum albumin (BSA) at a dilution of 1:500 and incubated at room temperature in a humid chamber for 1 hour. Coverslips were washed 9 times in PBS and then mounted onto glass slides using ProLong Gold antifade reagent with DAPI (#P36935, Thermofisher)

mounting media. Fluorescence microscopy was performed using a ZEISS CellDiscover 7 (CD7) microscope.

2.4.4. Human neutrophil infections

Blood was drawn from a healthy donor, using EDTA as an anticoagulant (BD Vacutainer, Becton Dickinson). Neutrophils were isolated using Polymorphprep (Serumwerk Bernburg) solution as per manufacturers guidelines. Briefly, 5-7 mL of blood was slowly layered over an equal quantity of Polymorphprep solution, and the layered solution was centrifuged for 30 minutes at 500 x *g* with decreased acceleration and no brakes on deceleration (to ensure separation of layers) (Fig 2.2). The neutrophil layer was collected and washed with 10 mL 0.5% (v/v) PBS. Where red blood cells (RBC) were present, a lysis step was incorporated, where 3 mL of 1X RBC lysis buffer (Invitrogen) was added per 5 mL of blood collected and incubated for 10 minutes. Following RBC lysis, neutrophils were washed twice as previously described and finally resuspended in neutrophil medium (RPMI 1640 Medium, GlutaMAX™ Supplement, Sigma Aldrich). Neutrophils were counted using Trypan Blue Staining and resuspended to the desired concentration.

Infections were performed in a 48-well plate (VWR), with 10^5 neutrophils added per well. *S. sonnei* was cultured as previously described and 10^3 bacteria were added per well, with experiments performed in technical duplicates. 20 µL of the initial inoculum was plated to determine precise bacterial input. The 48-well plate was incubated for 1 hour at 37 °C with 5% CO_2. For CFU determination, 7.5 µL of 0.1% Triton-X was added and neutrophils were placed on ice to lyse. A 10-fold serial dilution was performed and 20 µL of lysate at each dilution was plated. CFUs were counted the following morning and bacterial survival was determined by normalising CFU at 1 hpi to the initial inoculum.

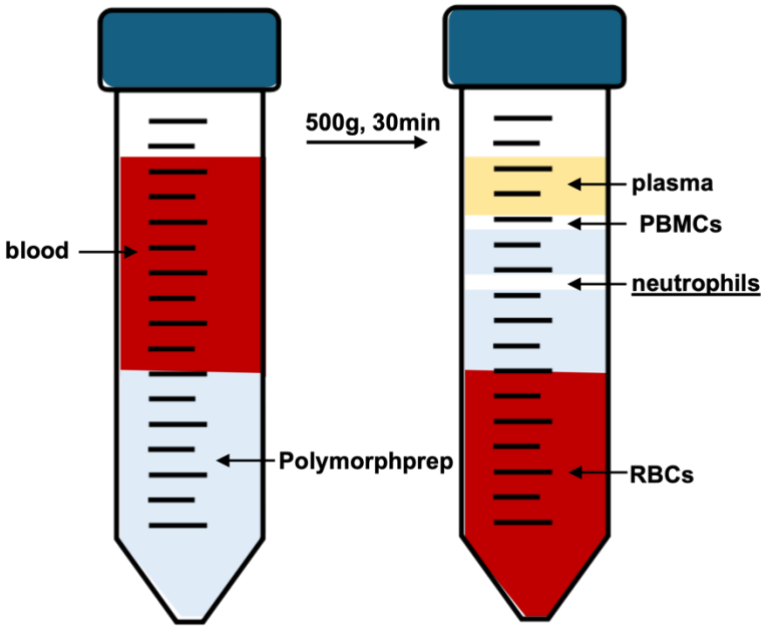


Figure 2.2. Neutrophil extraction using Polymorphprep. Schematic showing the procedure for isolating human neutrophils from whole blood. RBCs = red blood cells, PBMCs = peripheral blood mononuclear cells (PBMCs).

2.5. Bioinformatic methodologies

2.5.1. Phylogeny dating

Enterobase was used to identify publicly available ST99 *E. coli* genomes and all ST99 genomes with an associated assembly and isolation date were downloaded**.** Complete genome sequences of strains NCTC 9096 and CFSAN029787 were downloaded from GenBank (accessions UGEL01000000 and CP011416.1, respectively). To generate a core genome alignment, Snippy (v.4.6) was used and the complete genome of CFSAN029787 (ST99 EIEC) was used as a reference. Gubbins (v.3.2.1) was next used to identify recombinant regions of the resulting core SNP alignment and RaxML (v.8.10) (Stamatakis, 2014) was used to build a maximum likelihood phylogenetic tree, using the General Time Reversible (GTR) GAMMA nucleotide substitution model. To infer the dated phylogeny, BactDating (v.1.2) (Didelot et al., 2018) was used with the 'relaxedgamma' model, the option to incorporate Gubbins-detected recombination was selected, and 10^5 Markov chain Monte Carlo (MCMC) chain iterations were run, where convergence was achieved. To confirm the presence of temporal signal within the dataset, tip nodes were assigned random dates, and the analysis was re-run with the scrambled dates 10 times to confirm no overlap with real data and randomised dates.

2.5.2. Genome screening for pINV presence

To screen assemblies for the presence of pINV ShigEiFinder (v.1.3.4) (Zhang et al., 2021), for which the criteria is at least 26 of 38 pINV encoded genes present, was used. Tree visualisation and plotting of pINV presence was performed using Interactive Tree of Life (iTOL) (v.5.6.1) (Letunic and Bork, 2021).

2.5.3. De novo genome assembly

Prior to genomic analysis, sequencing reads were quality checked using FastQC (v.0.12.0) (Andrews, 2010), trimmed using Filtlong (v.0.2.1) (Wick, 2017) (for long reads) and Trimmomatic (v.0.4.0) (Bolger et al., 2014) (for short reads) using default settings. Genomes were then assembled using the Hybracter (v.0.7.3) (Bouras et al., 2024b) long-read first

assembly pipeline, with Flye (v.2.9.4) (Kolmogorov et al., 2019) selected as the long read assembler. As part of the pipeline, genomes were polished with Medaka (v.1.8.0) first, then with short reads using PyPOLCA (v.0.3.1) (Bouras et al., 2024c, Zimin and Salzberg, 2020). The quality of assemblies was checked using Quast (v.5.0.2) (Gurevich et al., 2013) (using Ss046 as a reference genome, accession: NC_007384) and completeness and contamination was analysed using CheckM (v.1.1.6) (Parks et al., 2015). Contigs were concatenated into a multifasta file using SeqKit (v.2.8.2) (Shen et al., 2016) prior to annotation and then genomes were annotated using Bakta (v.1.9.3) (Schwengers et al., 2021), with the full database option selected.

2.5.4. Genome characterisation

Complete genomes were genotyped using Mykrobe (v.0.13.0) (Hunt et al., 2019, Hawkey et al., 2021) to confirm *S. sonnei* lineage assignments obtained previously (Hawkey et al., 2021). Multilocus sequence typing (MLST) was performed using the Achtman 7 locus scheme hosted on pubMLST (accessed on 18th July 2024) (Jolley et al., 2018). Plasmids were replicon typed using MOBsuite (v.3.1.9) (Robertson and Nash, 2018) 'Mob-typer' function. Virulence factors were identified using ABRicate (v.1.0.1) (Seemann, 2016), with the VFDB option selected (accessed on 18/08/2024) (Liu et al., 2022) and AMR determinants were identified using NCBI AMRFinder (v.3.12.8) (Feldgarden et al., 2021). Colicins were identified using ABRicate (v.1.0.1) using the custom colicin database created by De Silva *et al* (De Silva et al., 2023) (https://figshare.com/articles/dataset/colicin_database/20768260/1?file=37009930).

Identification of IS elements was performed using ISEScan (v.1.7.2.3) (Xie and Tang, 2017) and results were filtered to identify ISs present in the chromosome, pINV and other plasmids. Bakta did not efficiently annotate pseudogenes (tested on 53G for which the number of pseudogenes is known), so the prokaryotic genome annotation pipeline (PGAP) (v.6.7) (Tatusova et al., 2016) was used to identify pseudogenes instead.

2.5.5. Pangenome analysis

Pangenome analysis was performed using Panaroo (v.1.5.0) (Tonkin-Hill et al., 2020) using the 'strict' mode. MAFFT (v. 7.526) (Katoh and Standley, 2013) was selected to perform a core genome alignment, and otherwise default settings were used. The phylogenetic tree for pangenome visualisation was produced using FastTree (v.2.1.1) (Price et al., 2010) using the generalised time reversible model and otherwise default settings. Visualisation was performed using Phandango (accessed on 18/08/2024) (Hadfield et al., 2017). Unique HGCs for each lineage were classified into biological process Gene Ontology (GO) categories using the PANTHER functional classification system (v.19.0) (Thomas et al., 2022).

2.5.6. Whole genome alignment

Where necessary, complete genome sequences were reoriented to the *dnaA* gene (for chromosomal sequences) and *repA* (for plasmid sequences) using dnaapler (v.0.7.0) (Bouras et al., 2024a) to ensure all genomes started from the same point. For lineage 3.7 isolates, sequences were reverse complemented before genome alignment, due to an inverted colinear block involving the *dnaA* gene and all genomes were reoriented to *fabB* where alignments including lineage 3.7 isolates were performed. Multiple whole genome alignments were performed using progressive Mauve (v.2.4.0) (Darling et al., 2010), with GenBank files generated by Bakta used as an input.

2.5.7 Gene cluster alignment

To generate gene cluster comparison figures, relevant gene clusters were extracted following visualisation in Mauve, aligned using Clinker (v.0.0.29) (Gilchrist and Chooi, 2021) and the output html file was visualised in Google Chrome.

2.6. Statistical analysis

Statistical analysis was performed in GraphPad Prism (v.9.4.1). Unless stated otherwise, data represents the mean ± standard error of the mean (SEM) from at least three independent biological replicates. For zebrafish survival curves, statistical significance was determined using the log-rank Mantel Cox test. In all other cases, significance was determined using an unpaired two-tailed Student's t test (for the comparison of two groups), a one-way ANOVA (where there are three or more independent groups) or a two-way ANOVA (where there were multiple independent variables) as indicated in figure legends. Where appropriate, Tukey's post-hoc test was applied to correct for multiple comparisons.

Chapter 3. Acquisition of a large virulence plasmid (pINV) promoted temperature-dependent virulence and global dispersal of O96:H19 enteroinvasive *Escherichia coli*

3.1. Introduction

3.1.1. Enteroinvasive *E. coli* (EIEC)

EIEC was first described in 1947, approximately 50 years after the discovery of *Shigella* (Ewing and Gravatti, 1947). EIEC belongs to the same diarrheagenic *E. coli* pathotype as *Shigella*, causing pINV-mediated shigellosis in the same way, but was not designated as *Shigella* due to discrete differences in virulence, biochemical and phenotypic characteristics. Genomic interrogation has revealed the presence of at least three EIEC lineages (which are distinct from *Shigella* lineages), with the presence of several other outlier strains that do not fall into the defined EIEC lineages, but other *E. coli* or *Shigella* phylogroups (Hazen et al., 2016). This study also revealed that like *Shigella*, EIEC is polyphyletic, with independent emergences from different *E. coli* ancestors following pINV acquisition occurring for each separate lineage. EIEC are commonly able to use a wider range of metabolites than *Shigella* subgroups (Hawkey et al., 2020) and have not undergone genome reduction to the same extent that *Shigella* has, with some of the known *Shigella* antivirulence loci still intact in EIEC lineages (Prosseda et al., 2012) (Table 3.1).

The global abundance of EIEC has been estimated in some regions but generally remains unclear (and likely underestimated) due to difficulties in distinguishing EIEC from both *Shigella* and other *E. coli* pathotypes (Gomes et al., 2016). EIEC reportedly causes somewhat less severe shigellosis than *Shigella* (DuPont et al., 1971) and in line with this, also exhibits a reduced expression of virulence genes encoding for T3SS components and effectors (Moreno et al., 2009). Together, the reduced virulence induced by EIEC infection, along with the biochemical and genomic similarities to commensal *E. coli* have led to the proposal that EIEC

represents an evolutionary intermediate between commensal *E. coli* and *Shigella*, with EIEC remaining in the earlier stages of host specialisation (Peng et al., 2009).

Table 3.1. Presence or absence of known antivirulence genes in *Shigella* and *E. coli.* ST = sequence type, EIEC = Enteroinvasive *E. coli*.

			Antivirulence loci			
Subgroup	**Strain**	**pINV**	*cad*	*nad*	*speG*	*ompT*
***E. coli* (commensal)**	MG1655 (K12)	-	+	+	+	+
ST6 EIEC	NCTC 10959	+	-	+	+	-
ST99 EIEC	152661	+	+	+	+	-
ST270 EIEC	8-3-Ti3	+	-	-	+	-
S. sonnei	53G	+	-	-	-	-
S. flexneri	M90T	+	-	-	-	-
S. boydii	NCTC 12985	+	-	-	-	-
S. dysenteriae	NCTC 9718	+	-	-	-	-

3.1.2. An emerging EIEC clone (O96:H19)

In 2012, a novel clone of EIEC was identified following an outbreak of foodborne diarrhoea in Italy, in which several people were hospitalised (Escher et al., 2014). Following this, a second large-scale outbreak in the United Kingdom was detected in 2014, and since then, the implicated clone has been circulating in Europe and South America (Newitt et al., 2016, Lagerqvist et al., 2020, Peirano et al., 2018). Bacteria isolated from the outbreaks were found to be of sequence type (ST)99 and serotype O96:H19, a serotype never associated with EIEC or *Shigella* before. Further characterisation found that O96:H19 EIEC harboured all the

classical virulence factors associated with *Shigella*, but its biochemical capabilities more closely resembled non-pathogenic *E. coli* than other EIEC lineages (Michelacci et al., 2016). Intriguingly, just one of the known *Shigella* and EIEC antivirulence loci (*ompT*) was disrupted in the genomes of O96:H19 EIEC, all isolates tested were motile and all had the conjugative *tra* loci of pINV intact (whilst it is disrupted in most other *Shigella* and EIEC subgroups), suggesting that O96:H19 had arisen as the result of a recent pINV acquisition.

3.1.3. Aims

This chapter exploits the recently emerged clone of EIEC, O96:H19, as a model system to explore the initial stages of *Shigella* and EIEC emergence. The specific aims of this chapter were to reconstruct the evolutionary history of ST99 EIEC and estimate a date for its emergence; to investigate differences in pathogenicity between EIEC and non-EIEC ST99 isolates; and to explore the role of pINV acquisition in driving the virulence of ST99 EIEC.

3.2.1 Publication: Acquisition of a large virulence plasmid (pINV) promoted temperature-dependent virulence and global dispersal of O96:H19 enteroinvasive *Escherichia coli*

RESEARCH PAPER COVER SHEET

Please note that a cover sheet must be completed <u>for each</u> research paper included within a thesis.

SECTION A – Student Details

Student ID Number	LSH2006138	**Title**	Miss
First Name(s)	Sydney-Leigh		
Surname/Family Name	Miles		
Thesis Title	Using zebrafish to study the evolution and pathogenesis of *Shigella*		
Primary Supervisor	Serge Mostowy		

If the Research Paper has previously been published please complete Section B, if not please move to Section C.

<u>SECTION B – Paper already published</u>

Where was the work published?	mBio		
When was the work published?	May 2023		
If the work was published prior to registration for your research degree, give a brief rationale for its inclusion	NA		
Have you retained the copyright for the work?*	**Yes**	Was the work subject to academic peer review?	**Yes**

*If yes, please attach evidence of retention. If no, or if the work is being included in its published format, please attach evidence of permission from the copyright holder (publisher or other author) to include this work.

SECTION C – Prepared for publication, but not yet published

Where is the work intended to be published?	
Please list the paper's authors in the intended authorship order:	
Stage of publication	Choose an item.

SECTION D – Multi-authored work

For multi-authored work, give full details of your role in the research included in the paper and in the preparation of the paper. (Attach a further sheet if necessary)	The article was written by me, Sydney Miles, with guidance from supervisors, Prof. Serge Mostowy and Prof. Kathryn Holt. All experiments and bioinformatic analysis were performed by me, except for the SDS-PAGE blot which was performed with the help of Dr Ana Teresa López-Jiménez. Dr Zoe Dyson and Dr Ebenezer Foster Nyarko provided guidance on bioinformatic analysis. Dr Damián Lobato-Márquez provided guidance on cloning.

	Dr Claire Jenkins provided access to clinical isolates. Dr Vincenzo Torraca, Prof Serge Mostowy and Prof Kathryn Holt provided formal supervision and conceptualised the project.

SECTION E

Student Signature	Sydney-Leigh Miles
Date	04/09/2024

Supervisor Signature	Serge Mostowy
Date	04/09/2024

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

Check for updates

| Bacteriology | Observation

Acquisition of a large virulence plasmid (pINV) promoted temperature-dependent virulence and global dispersal of O96:H19 enteroinvasive *Escherichia coli*

Sydney L. Miles,[1] Vincenzo Torraca,[1] Zoe A. Dyson,[1,2,3] Ana Teresa López-Jiménez,[1] Ebenezer Foster-Nyarko,[1] Damián Lobato-Márquez,[1] Claire Jenkins,[4] Kathryn E. Holt,[1,2] Serge Mostowy[1]

AUTHOR AFFILIATIONS See affiliation list on p. 8.

ABSTRACT Enteroinvasive *Escherichia coli* (EIEC) and *Shigella* are closely related agents of bacillary dysentery. It is widely viewed that EIEC and *Shigella* species evolved from *E. coli* via independent acquisitions of a large virulence plasmid (pINV) encoding a type 3 secretion system (T3SS). Sequence Type (ST)99 O96:H19 *E. coli* is a novel clone of EIEC responsible for recent outbreaks in Europe and South America. Here, we use 92 whole genome sequences to reconstruct a dated phylogeny of ST99 *E. coli*, revealing distinct phylogenomic clusters of pINV-positive and -negative isolates. To study the impact of pINV acquisition on the virulence of this clone, we developed an EIEC-zebrafish infection model showing that virulence of ST99 EIEC is thermoregulated. Strikingly, zebrafish infection using a T3SS-deficient ST99 EIEC strain and the oldest available pINV-negative isolate reveals a separate, temperature-independent mechanism of virulence, indicating that ST99 non-EIEC strains were virulent before pINV acquisition. Taken together, these results suggest that an already pathogenic *E. coli* acquired pINV and that virulence of ST99 isolates became thermoregulated once pINV was acquired.

IMPORTANCE Enteroinvasive *Escherichia coli* (EIEC) and *Shigella* are etiological agents of bacillary dysentery. Sequence Type (ST)99 is a clone of EIEC hypothesized to cause human disease by the recent acquisition of pINV, a large plasmid encoding a type 3 secretion system (T3SS) that confers the ability to invade human cells. Using Bayesian analysis and zebrafish larvae infection, we show that the virulence of ST99 EIEC isolates is highly dependent on temperature, while T3SS-deficient isolates encode a separate temperature-independent mechanism of virulence. These results indicate that ST99 non-EIEC isolates may have been virulent before pINV acquisition and highlight an important role of pINV acquisition in the dispersal of ST99 EIEC in humans, allowing wider dissemination across Europe and South America.

KEYWORDS EIEC, zebrafish, host-pathogen interactions, evolution, *Shigella*, *Enterobacteriaceae*, virulence determinants

Enteroinvasive *E. coli* (EIEC) and *Shigella* species are Gram-negative, human-adapted pathogens that cause bacillary dysentery. The greatest burden of bacillary dysentery is in low- and middle-income countries (LMICs) (1), although the true burden of EIEC infection is likely underestimated since it is difficult to distinguish from *Shigella*. Historically, *Shigella* was classified as its own genus, with four distinct species, but Multi-Locus Sequence Typing (MLST) and whole-genome sequencing data clearly show *Shigella* spp. are lineages of *E. coli*, as are EIEC (2, 3). Each *Shigella* and EIEC lineage evolved independently within the *E. coli* population, following the horizontal acquisition of a ~220 kbp virulence plasmid (also known as plasmid of invasion or pINV) from a

Editor Carmen Buchrieser, Institut Pasteur, Paris, France

Address correspondence to Kathryn E. Holt, kat.holt@lshtm.ac.uk, or Serge Mostowy, serge.mostowy@lshtm.ac.uk.

The authors declare no conflict of interest.

See the funding table on p. 9.

Received 10 April 2023
Accepted 17 April 2023
Published 31 May 2023

currently unknown source (2). pINV encodes a type three secretion system (T3SS) that facilitates the invasion of human epithelial cells and is thermoregulated in both EIEC and *Shigella* (4).

A novel clone of EIEC, of serotype O96:H19 and Multi-Locus Sequence Type (ST) 99, was first described in 2012 in Italy and has since caused several foodborne outbreaks of moderate to severe diarrheal disease across Europe and South America (5–7). Before 2012, ST99 *E. coli* had not been reported in the literature as causing human disease but had been sporadically isolated from cattle and environmental sources (8). ST99 EIEC isolates have been characterized as possessing the virulence hallmarks of EIEC and *Shigella* (pINV and T3SS) (9), but its metabolic capacity closely resembles that of commensal *E. coli* and it has more recently been associated with *pga*-mediated biofilm formation (6, 9). It has therefore been proposed that ST99 EIEC diverged recently from ST99 *E. coli* due to the acquisition of pINV.

The zebrafish (*Danio rerio*) larvae model is widely used to study infection biology *in vivo* because of its rapid development and innate immune system that is highly homologous to that of humans (10, 11). Zebrafish have emerged as a valuable vertebrate model to study human enteropathogens like *Shigella* (12), highlighting the key roles of bacterial virulence factors (e.g., T3SS and O-antigen) (13, 14) and cell-autonomous immunity (e.g., autophagy and septin-mediated immunity) (12, 15) in host-pathogen interactions.

In this observation, we reconstruct a dated phylogeny of ST99 *E. coli* using publicly available whole genome sequences, to understand the role of pINV in its global dispersal. We develop a temperature-dependent zebrafish infection model to assess the virulence of EIEC and non-EIEC ST99 isolates, highlighting the power of zebrafish infection in studying the evolution of novel enteropathogens causing disease in humans.

ST99 EIEC diverged ~40 years ago

To dissect the evolution of the ST99 clone and its transition to EIEC, we analyzed all publicly available ST99 genomes (n = 92), using the EnteroBase integrated software environment (16). EnteroBase routinely scans short-read archives and retrieves *E. coli* and *Shigella* sequences from the public domain or uses user-uploaded short reads. We used Gubbins v.3.2.1 (17) to filter recombinant sites, RaxML v.8.10 to infer a Maximum Likelihood phylogenetic tree and BactDating v.1.2 (18) to date the phylogeny (Fig. 1), as previously described by Didelot and Parkhill (19). Root-to-tip genetic distances were positively associated with the year of isolation (R^2 = 0.19, $P = 6 \times 10^{-3}$), and the date-randomization test showed no overlap between results of observed and date-randomized analyses (Fig. S1), indicating a moderate molecular clock signal to support dating analysis. From this analysis, we estimate that the most recent common ancestor (MRCA) of the whole ST99 group (pINV+ and pINV–) existed circa 1776 [95% highest posterior density (HPD), 1360–1927]. To test for the presence of pINV, we used ShigEiFinder, which scans the genomes for pINV-encoded genes and deems an isolate positive for pINV when 26 of 38 genes are present (20). The pINV+ isolates form a distinct cluster, with their MRCA existing circa 1982 (95% HPD, 1965–2011) (Fig. 1). This suggests that the ST99 EIEC may have been circulating undetected for ~30 years before being detected in the 2012 outbreak.

To test the role of pINV in the dispersal of ST99 EIEC, we selected: (i) four recent pINV+ isolates from moderate-to-severe diarrheal outbreaks in the United Kingdom in 2014 and 2015 (21, 22) to represent contemporary ST99 EIEC, (ii) a Congo red-negative colony to represent a T3SS-deficient strain isogenic strain, and (iii) the oldest available ST99 isolate (~1945, NCTC 9096, pINV–) to represent ancestral pINV– ST99 (see Fig. 1).

ST99 EIEC virulence is temperature-dependent in zebrafish

The zebrafish infection model has generated fundamental advances in our understanding of *Shigella* and its ability to infect humans (23). To test the virulence of pINV+ ST99 EIEC strains, ~5,000 CFU was injected into the hindbrain ventricle (HBV) of zebrafish

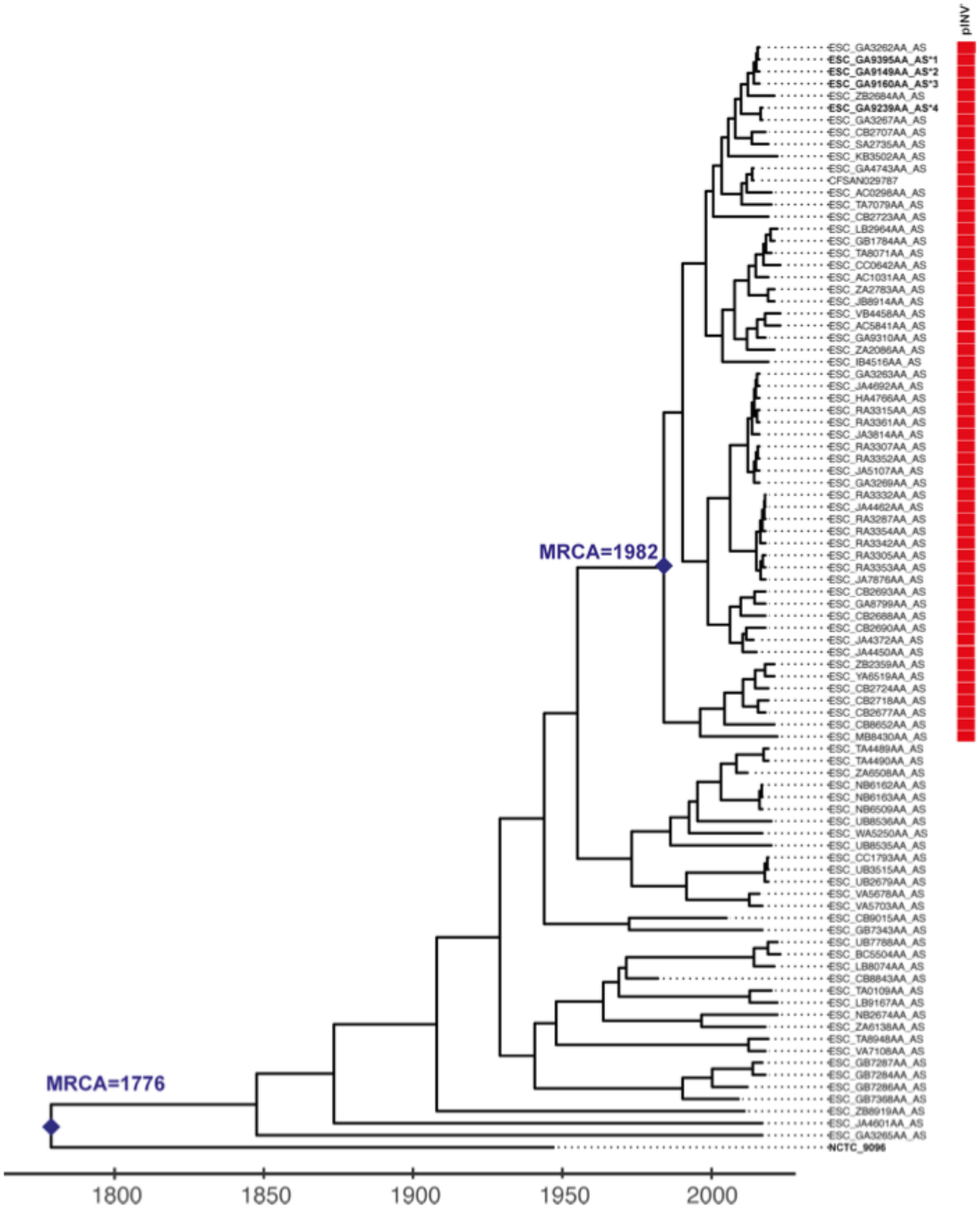


FIG 1 Time-calibrated phylogeny of 92 Sequence Type (ST)99 genomes. BactDating was used to infer a time-calibrated phylogeny, incorporating the output from the recombination detection software, Gubbins. Blue diamonds indicate the internal nodes representing the most recent common ancestors (MRCA) of interest. Tip labels represent assembly barcodes correlating to the isolate accession in Enterobase. Tip labels in bold represent isolates that we tested *in vivo*. As determined using ShigEiFinder (20), the presence of the invasion plasmid (pINV) is indicated by a red box in the pINV column. We estimate the MRCA of the whole group to be ~1776 and the MRCA for the pINV+ cluster to be ~1982.

larvae at 3 d post-fertilization (dpf) (Fig. S1A). Infected zebrafish larvae are typically incubated at 28.5°C for optimal development but we have shown they can also be maintained at 32.5°C (13), allowing the study of temperature-dependent virulence. For the pINV+ strains, we observed ~75% survival when larvae were incubated at 28.5°C but only ~30% survival when incubated at 32.5°C (Fig. 2A; Fig. S2B and C). In agreement with survival results, CFUs recovered at 6 h post-infection (hpi) were significantly lower at 28.5°C than CFUs recovered at 32.5°C (Fig. 2B; Fig. S2D and E), suggesting that larvae were more able to control infection at 28.5°C.

To test if the T3SS in ST99 EIEC is functional and thermoregulated, we compared the secretion of virulence factors by ST99 EIEC and *Shigella flexneri in vitro* (Fig. S3). The overall abundance of secreted proteins is lower for ST99 EIEC as compared to *S. flexneri*, but the relative abundance of major secreted effectors appears similar. One exception is SepA, a protein secreted independently of the T3SS, whose presence is known to be variable in other EIEC lineages (24). Although we do not observe significant differences in secretion between 28.5°C and 32.5°C under these *in vitro* conditions tested, the T3SS in ST99 EIEC is clearly thermoregulated (with optimal secretion *in vitro* at 37°C).

Having established a temperature-dependent EIEC-zebrafish infection model, it was next of great interest to test the virulence of an isogenic, T3SS-deficient ST99 EIEC strain and an ancestral pINV– ST99 isolate. Since we observed no significant differences in zebrafish survival or bacterial burden between the four pINV+ strains at either 28.5°C or 32.5°C (Fig. S2B to E), we chose one isolate (pINV+1) as a representative pINV+ isolate to compare with the T3SS-deficient and pINV– isolate (NCTC 9096).

ST99 *E. coli* comprises temperature-dependent and -independent mechanisms of virulence

To test if the virulence of ST99 *E. coli* in zebrafish is dependent on the acquisition of pINV and the T3SS, we selected a naturally T3SS-deficient colony (Congo red negative) to compare against pINV+1. Colonies were screened for several pINV-encoded genes by colony PCR and found to be deficient in genes located in the T3SS-encoding region of pINV (*mxiG*, *mxiD,* and *icsB*), but positive for genes located outside (*ospF* and *ipaH*) (Fig. S4). In addition, we verified dysfunction of the T3SS, showing that T3SS effector proteins are not secreted by T3SS-deficient colonies at 37°C but are by the wild-type EIEC isolate (Fig. S3). These results suggest that the T3SS-encoding region has been lost in Congo red negative colonies, consistent with what has previously been reported for *S. flexneri* (25). Infection of zebrafish with these isolates shows that the thermoregulated virulence is lost in the T3SS-deficient strain, with no significant difference in zebrafish survival observed between 28.5°C and 32.5°C, whilst thermoregulated virulence is maintained in the wild-type strain (Fig. 2C and D; Fig. S5A and B). These result implicate acquisition of the T3SS (and pINV) in the temperature-dependent virulence of pINV+1.

Next, we compared the virulence of a non-EIEC (pINV–) ST99 isolate (NCTC 9096) and a pINV+ EIEC isolate (pINV+1) strains using our EIEC-zebrafish infection model. Strikingly, NCTC 9096 was significantly more virulent than pINV+1 at 28.5°C, with only ~35% of infected larvae surviving at 48 hpi (Fig. 2E). Although no change in survival of larvae infected with NCTC 9096 is observed at 32.5°C (as compared to that of 28.5°C), survival of pINV+1 infected larvae significantly decrease at 32.5°C, consistent with a role for temperature-dependent virulence. At 32.5°C, we found that both pINV+1 and NCTC 9096 isolates were equally virulent (Fig. 2F).

The trend in virulence was also reflected in the quantification of bacterial burden (Fig. S5C and D). When incubated at 32.5°C, we observed a ~2 log increase in pINV+1 CFUs enumerated from larvae at 6 hpi but not when incubated at 28.5°C. We observe a similar increase in NCTC 9096 CFUs quantified, irrespective of temperature. These results show temperature-dependent virulence of the pINV+1 strain and non-temperature-dependent virulence of the pINV– strain, consistent with our observations for the T3SS-deficient EIEC isolate.

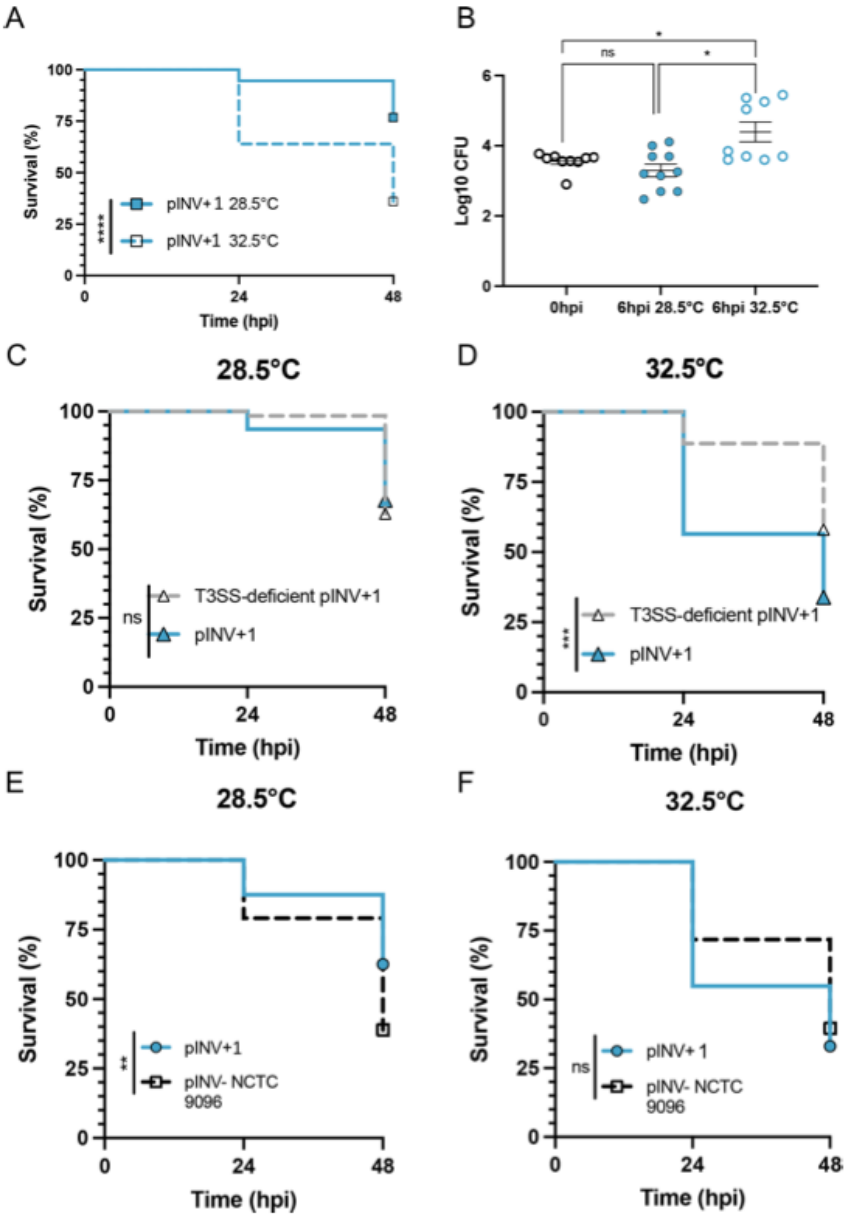


FIG 2 Temperature-dependent and -independent mechanisms of virulence in the ST99 group. Zebrafish larvae at 3 d post-fertilization were injected with 5,000 CFU of a representative pINV+ ST99 strain, a T3SS-deficient strain and an ancestral pINV− ST99 strain, before being separated for incubation at 28.5°C or 32.5°C. (**A, B**) pINV+1 strain exhibits a temperature-dependent virulence with significantly more killing observed at 32.5°C. Enumeration of bacterial burden is also temperature dependent, with greater CFUs quantified at 6 h post-infection from larvae incubated at 32.5°C. Black circles indicate pINV+1 CFUs at 0 hpi, blue filled circles indicate pINV+1 CFUs at 6 hpi incubated at 28.5°C, and blue outlined circles indicate pINV+1 CFUs at 6 hpi incubated at 32.5°C. (**C, D**) Thermoregulated virulence is lost in a T3SS-deficient (Congo red negative) pINV+1 strain (gray dashed line). (**E, F**) pINV− strain NCTC 9096 (black dashed line) is virulent in the zebrafish model in a non-temperature-dependent manner. Significance was tested using Log-rank (Mantel-Cox) test for survival curves. For CFUs (panel B), significance was tested using a one-way ANOVA with Sidak's correction. *$P < 0.0332$; **$P < 0.0021$; ***$P < 0.0002$; and ****$P < 0.0001$.

DISCUSSION

It is widely recognized that the acquisition of pINV is a defining feature in the evolution of EIEC and *Shigella* (26). Here, we analyze the evolution of ST99 EIEC and propose that an MRCA for the pINV+ group existed in the early 1980s. This suggests that ST99 EIEC may have been circulating undetected for ~30 years until it was implicated in the 2012 outbreak in Italy, perhaps because EIEC infections are typically endemic in regions where surveillance and sequencing of enteropathogens are limited.

We prove that the virulence of ST99 EIEC strains is thermoregulated *in vitro* and *in vivo* (with zebrafish larvae less able to control infection), leading to increased killing and greater bacterial replication at 32.5°C. Some killing is still observed at 28.5°C, suggesting a low-level activation of the T3SS and/or non-T3SS mechanisms of virulence *in vivo*, which would be of interest to test in future studies. These data are consistent with previous reports for pINV-mediated virulence in both *S. flexneri* and *Shigella sonnei* (4, 13). Our zebrafish infection model highlights the importance of temperature in EIEC virulence and supports the hypothesis that pINV acquisition is the first key step in the evolutionary pathway toward becoming a human-adapted pathogen. Our data using zebrafish infection further show that non-EIEC ST99 isolates can also cause disease and that the ability of the ST99 clone to cause disease does not strictly rely on the acquisition of pINV and the transition to EIEC. Considering that pINV– ST99 strain NCTC 9096 is highly virulent *in vivo*, we conclude that it must encode separate, non-thermoregulated mechanism(s) of virulence that becomes less important for human infection once pINV is acquired.

Collectively, our findings illuminate the short history of ST99 EIEC and implicate pINV acquisition as a key factor in its epidemiological success. Our approach also reveals a separate, non-thermoregulated virulence mechanism in a pINV– ST99 isolate, suggesting that an already pathogenic *E. coli* may have acquired pINV. Further studies, including identifying the source of pINV and those isolates likely to acquire it, are important to fully understand and prevent the dispersal of novel EIEC and *Shigella* clones infecting humans.

METHODS

Bacterial strains

Four pINV+ EIEC strains isolated in diarrhoeal outbreaks from the United Kingdom in 2014/2015 were included in this study (Table 1), and strains were identified and sequenced through routine surveillance and kindly shared with us by the UK Health and Security Agency (UKHSA). A pINV– ST99 strain (Table 1) included in this study was obtained from the National Culture Type Collection (NCTC). *S. flexneri* M90T was used as a positive control for the *in vitro* secretion assay (27).

TABLE 1 Bacterial strains used in experimental work/as a reference genome[a]

Strain	Source	Serotype	pINV	Origin	Sequence accession no.	Enterobase assembly barcode
NCTC 9096	NCTC	O96:H19	–	Denmark, 1945	UGEL01000000	ESC_CC4859AA_AS
pINV+1	UKHSA	O96:H19	+	United Kingdom (Travel to Turkey), 2014	SRR3578973	ESC_GA9395AA_AS
pINV+2	UKHSA	O96:H19	+	Kingdom (Travel to Turkey), 2014	SRR3578582	ESC_GA9149AA_AS
pINV+3	UKHSA	O96:H19	+	Kingdom (Travel to Turkey), 2014	SRR3578593	ESC_GA9160AA_AS
pINV+4	UKHSA	O96:H19	+	United Kingdom, 2015	SRR3578770	ESC_GA9239AA_AS
CFSAN029787	NA	O96:H19	+	Italy, 2012	Chromosome: CP011416.1, pINV: CP011417.1	ESC_GA4743AA_AS
S. flexneri M90T	Institut Pasteur	5 a	+	Mexico, 1955	NA[b]	NA[b]

[a]Strains NCTC 9096 and the four pINV+ ST99 strains obtained from the UKHSA were used for the *in vivo* work. CFSAN029787 was used as a reference strain for the phylogenomic analyses. Enterobase assembly accessions correlate to tip labels in the dated phylogeny (Fig. 1).
[b]NA, not applicable.

To obtain a T3SS-deficient ST99 isolate, bacteria (pINV+1) were grown on trypticase soy agar (TSA) plates supplemented with 0.01% Congo red (Sigma-Aldrich) dye. A white colony was selected as a natural isogenic mutant, unable to secrete T3SS effector proteins, as previously described (28). We tested this isolate for the presence of five pINV-encoded genes (*mxiG, mxiD, icsB, ipaH,* and *ospF*) by colony PCR, using primers described in Table 2.

Genomic analysis

Enterobase was used to identify publicly available ST99 genomes, using the filter by ST function (16); all ST99 genomes with an associated assembly and isolation date were included in our study, sequence accessions and metadata can be found in Table S1. Complete genome sequences of strains NCTC 9096 and CFSAN029787 were downloaded from GenBank (accessions UGEL01000000 and CP011416.1, respectively). All genomic analyses were performed using the Cloud Infrastructure for Microbial Bioinformatics (CLIMB) (29). Snippy v.4.6 (https://github.com/tseemann/snippy) was used to generate a core genome alignment, using CFSAN029787 as the reference. Gubbins v.3.2.1 (17) was used to identify recombinant regions of the alignment, and RaxML v.8.10 (30) was used to build a maximum likelihood phylogenetic tree, using the General Time Reversible (GTR) GAMMA nucleotide substitution model. BactDating v.1.2 (31) was used to infer the dated phylogeny, using the "relaxedgamma" model, the option to incorporate Gubbins detected recombination was selected and 10^5 Markov chain Monte Carlo (MCMC) chain iterations were run. To confirm the temporal signal (association between genetic divergence and time) within our dataset, tip nodes were assigned random dates and the analysis was rerun (this was completed $n = 10$ times). We saw no overlap between the substitution rates of our real data and the randomized datasets (Fig. S3) that shows that the data pass the stringent test CR2 for the presence of a temporal signal according to Duchene et al. (32). To screen assemblies for the presence of pINV, ShigEiFinder was used, which screens for 38 pINV-encoded genes and deems an isolate positive when at least 26 genes are present (20).

Inoculate preparation

Single red colonies (pINV+ EIEC) or white colonies (pINV– NCTC 9096 and T3SS-deficient EIEC) were selected and inoculated into 5 mL trypticase soy broth (TSB) and incubated overnight at 37°C, shaking at 400 rpm. 400 µL of overnight culture was subsequently diluted in 20 mL TSB and grown until an optical density of ~0.6 (measured at 600 nm) was reached. For zebrafish larvae infections, inoculate preparation was carried out by resuspension of the bacteria at the desired concentration in phosphate buffer saline (PBS, Sigma-Aldrich) pH 7.4 containing 2% polyvinylpyrrolidone (Sigma-Aldrich) and 0.5% phenol red (Sigma-Aldrich) as previously described (13).

TABLE 2 Primers used to detect for the presence of pINV-encoded genes

Primer name	Primer sequence (5'-3')
mxiD_Fwd	CAGAATGTAAGTAATGCACTGGCTATGATAC
mxiD_Rev	CTGTCTATAAAATCCTGATCTAGAGGAAGGTTATC
mxiG_Fwd	CTGATTGTTGGGATAAGGCTGG
mxiG_Rev	CCGAGATCCCCTGTTTACCTC
ospF_Fwd	AAAAGATGAAGGCCTGATGGGAGCATTAAC
ospF_Rev	TGGTGGATAAAACCCGCCAGAATGAACA
icsB_Fwd	GGTTCCAAGATCTGGCGATTTAAGAGAATTGTAATAATC
icsB_Rev	GGGCCTATACGCGTTGAAGATACAGAG
ipaH1.4_Fwd	GGGCATGAAAAAAGCTACATCC
ipaH1.4_Rev	CACCATTATTCGAGTATAGGGAGAG

Zebrafish larvae infection

Wild-type-AB zebrafish embryos were used for *in vivo* studies. Embryos were kept in 0.5 × E2 medium supplemented with 0.3 µg/mL methylene blue and incubated at 28.5°C unless otherwise stated. Using a microinjector, ~1 nL of bacterial suspension was injected into the HBV of 3 d post-fertilization (dpf) zebrafish larvae, following previously described procedures (13). The precise inoculum was determined retrospectively by homogenization of larvae at 0 h post-infection and plating on TSA plates supplemented with 0.01% Congo red.

For survival assays, zebrafish larvae were visualized using a light stereomicroscope at 24 and 48 hpi; the presence of a heartbeat was used to determine viability. For colony forming unit (CFU) counts, larvae were disrupted in PBS using a pestle pellet blender at 0 and 6 hpi. Serial dilutions in PBS and plating on TSA plates supplemented with 0.01% Congo red were then performed to estimate the bacterial load in each larva. Statistical analysis was performed in GraphPad Prism 9.

In vitro secretion assay

Secretion of T3SS effectors was tested as previously described (33). Briefly, bacteria were grown overnight, subcultured and grown until exponential phase (OD = 0.4–0.5) at either 28.5°C, 32.5°C, or 37°C. Cultures were then incubated for 3 h in the presence or absence of Congo red to induce type 3 secretion. Secreted proteins were collected from culture supernatants, precipitated using trichloroacetic acid (Sigma-Aldrich), and then analyzed using SDS-PAGE and Coomassie Brilliant Blue R-250 (Bio-Rad) staining.

ACKNOWLEDGMENTS

We thank members of the Mostowy and Holt labs for the helpful discussion. We thank the Biological Services Facility (BSF) at LSHTM for assistance with zebrafish care and breeding.

S.L.M. is supported by a Biotechnology and Biological Sciences Research Council LIDo Ph.D. studentship (BB/T008709/1). V.T. was supported by an LSHTM/Wellcome Trust Institutional Strategic Support Fund (ISSF) Fellowship (204928/Z/16/Z). A.T.L.J. is funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska – Curie Individual Fellowship (grant agreement No. H2020-MSCA-IF2020-895330). Work in the S.M. laboratory is supported by a European Research Council Consolidator Grant (grant agreement No. 772853-ENTRAPMENT) and Wellcome Trust Senior Research Fellowship (206444/Z/17/Z).

AUTHOR AFFILIATIONS

[1]Department of Infection Biology, London School of Hygiene and Tropical Medicine, London, United Kingdom
[2]Department of Infectious Diseases, Central Clinical School, Monash University, Melbourne, Victoria, Australia
[3]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, United Kingdom
[4]Gastrointestinal Pathogens and Food Safety (One Health), UK Health Security Agency, London, United Kingdom

PRESENT ADDRESS

Vincenzo Torraca, School of Life Sciences, University of Westminster, London, United Kingdom
Damián Lobato-Márquez, National Center of Biotechnology CSIC, Madrid, Spain

AUTHOR ORCIDs

Sydney L. Miles http://orcid.org/0000-0003-2291-4105

Vincenzo Torraca http://orcid.org/0000-0001-7340-0249
Zoe A. Dyson http://orcid.org/0000-0002-8887-3492
Ana Teresa López-Jiménez http://orcid.org/0000-0002-0289-738X
Ebenezer Foster-Nyarko http://orcid.org/0000-0001-6620-9403
Damián Lobato-Márquez http://orcid.org/0000-0002-0044-7588
Kathryn E. Holt http://orcid.org/0000-0003-3949-2471
Serge Mostowy http://orcid.org/0000-0002-7286-6503

FUNDING

Funder	Grant(s)	Author(s)
UKRI | Biotechnology and Biological Sciences Research Council (BBSRC)	BB/T008709/1	Sydney Leigh Miles
London School of Hygiene and Tropical Medicine (LSHTM)	204928/Z/16/Z	Vincenzo Torraca
EC | H2020 | PRIORITY 'Excellent science' | H2020 Marie Skłodowska-Curie Actions (MSCA)	H2020-MSCA-IF2020-895330	Ana Teresa López-Jiménez
EC | H2020 | PRIORITY 'Excellent science' | H2020 European Research Council (ERC)	772853-ENTRAPMENT	Serge Mostowy
Wellcome Trust (WT)	206444/Z/17/Z	Serge Mostowy

AUTHOR CONTRIBUTIONS

Sydney L. Miles, Conceptualization, Formal analysis, Funding acquisition, Investigation, Writing – original draft, Writing – review and editing | Vincenzo Torraca, Supervision, Writing – review and editing | Zoe A. Dyson, Formal analysis, Writing – review and editing | Ana Teresa López-Jiménez, Investigation, Writing – review and editing | Ebenezer Foster-Nyarko, Formal analysis, Writing – review and editing | Damián Lobato-Márquez, Investigation, Writing – review and editing | Claire Jenkins, Resources, Writing – review and editing | Kathryn E. Holt, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review and editing | Serge Mostowy, Conceptualization, Funding acquisition, Project administration, Supervision, Writing – original draft, Writing – review and editing

DATA AVAILABILITY STATEMENT

All genomes used in this study are publicly available in Enterobase, and assembly accessions are provided in Table S1.

ETHICS APPROVAL

Animal experiments were performed according to the Animals (Scientific Procedures) Act 1986 and approved by the Home Office (Project license: P4E664E3C). All experiments were conducted up to 5 days post-fertilization (dpf).

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Fig S1 (mBio00882-23-s0001.pdf). Results of date randomization test.
Fig S2 (mBio00882-23-s0002.pdf). Virulence of ST99 EIEC isolates is temperature dependent.

Fig S3 (mBio00882-23-s0003.pdf). Secretion of virulence factors *in vitro* by ST99 EIEC pINV+1 (Congo red+ colony), T3SS-deficient ST99 EIEC (Congo red- colony), and *S. flexneri*.
Fig S4 (mBio00882-23-s0004.pdf). Colony PCR to check for the presence of pINV-encoded genes.
Fig S5 (mBio00882-23-s0005.pdf). Bacterial burden of pINV+1 compared to its T3SS-deficient counterpart and the oldest available pINV– isolate, NCTC 9096.
TABLE S1 (mBio00882-23-s0006.xlsx). Details and metadata for all genomes downloaded from Enterobase.

REFERENCES

1. Kotloff KL, Nataro JP, Blackwelder WC, Nasrin D, Farag TH, Panchalingam S, Wu Y, Sow SO, Sur D, Breiman RF, Faruque AS, Zaidi AK, Saha D, Alonso PL, Tamboura B, Sanogo D, Onwuchekwa U, Manna B, Ramamurthy T, Kanungo S, Ochieng JB, Omore R, Oundo JO, Hossain A, Das SK, Ahmed S, Qureshi S, Quadri F, Adegbola RA, Antonio M, Hossain MJ, Akinsola A, Mandomando I, Nhampossa T, Acácio S, Biswas K, O'Reilly CE, Mintz ED, Berkeley LY, Muhsen K, Sommerfelt H, Robins-Browne RM, Levine MM. 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. Lancet 382:209–222. https://doi.org/10.1016/S0140-6736(13)60844-2
2. Pupo GM, Lan R, Reeves PR. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proc Natl Acad Sci U S A 97:10567–10572. https://doi.org/10.1073/pnas.180094797
3. Sahl JW, Morris CR, Emberger J, Fraser CM, Ochieng JB, Juma J, Fields B, Breiman RF, Gilmour M, Nataro JP, Rasko DA. 2015. Defining the phylogenomics of *Shigella* species: a pathway to diagnostics. J Clin Microbiol 53:951–960. https://doi.org/10.1128/JCM.03527-14
4. Falconi M, Colonna B, Prosseda G, Micheli G, Gualerzi CO. 1998. Thermoregulation of *Shigella* and *Escherichia coli* EIEC pathogenicity. a temperature-dependent structural transition of DNA modulates accessibility of virF promoter to transcriptional repressor H-NS. EMBO J 17:7033–7043. https://doi.org/10.1093/emboj/17.23.7033
5. Escher M, Scavia G, Morabito S, Tozzoli R, Maugliani A, Cantoni S, Fracchia S, Bettati A, Casa R, Gesu GP, Torresani E, Caprioli A. 2014. A severe foodborne outbreak of diarrhoea linked to a canteen in Italy caused by enteroinvasive *Escherichia coli*, an uncommon agent. Epidemiol Infect 142:2559–2566. https://doi.org/10.1017/-S0950268814000181
6. Iqbal J, Malviya N, Gaddy JA, Zhang C, Seier AJ, Haley KP, Doster RS, Farfán-García AE, Gómez-Duarte OG. 2022. Enteroinvasive *Escherichia coli* O96:H19 is an emergent Biofilm-forming pathogen. J Bacteriol Res 204: e0056221. https://doi.org/10.1128/jb.00562-21
7. Newitt S, MacGregor V, Robbins V, Bayliss L, Chattaway MA, Dallman T, Ready D, Aird H, Puleston R, Hawker J. 2016. Two linked enteroinvasive *Escherichia coli* outbreaks, Nottingham, UK, June 2014. Emerg Infect Dis 22:1178–1184. https://doi.org/10.3201/eid2207.152080
8. Bai X, Scheutz F, Dahlgren HM, Hedenström I, Jernberg C. 2021. Characterization of clinical *Escherichia* coli strains producing a novel shiga toxin 2 subtype in Sweden and Denmark. Microorganisms 9:2374. https://doi.org/10.3390/microorganisms9112374
9. Michelacci V, Prosseda G, Maugliani A, Tozzoli R, Sanchez S, Herrera-León S, Dallman T, Jenkins C, Caprioli A, Morabito S. 2016. Characterization of an emergent clone of enteroinvasive *Escherichia* coli circulating in Europe. Clin Microbiol Infect 22:287. https://doi.org/10.1016/j.cmi.2015.10.025
10. Howe K, Clark MD, Torroja CF, Torrance J, Berthelot C, Muffato M, Collins JE, Humphray S, McLaren K, Matthews L, McLaren S, Sealy I, Caccamo M, Churcher C, Scott C, Barrett JC, Koch R, Rauch G-J, White S, Chow W, Kilian B, Quintais LT, Guerra-Assunção JA, Zhou Y, Gu Y, Yen J, Vogel J-H, Eyre T, Redmond S, Banerjee R, Chi J, Fu B, Langley E, Maguire SF, Laird GK, Lloyd D, Kenyon E, Donaldson S, Sehra H, Almeida-King J, Loveland J, Trevanion S, Jones M, Quail M, Willey D, Hunt A, Burton J, Sims S, McLay K, Plumb B, Davis J, Clee C, Oliver K, Clark R, Riddle C, Elliot D, Threadgold G, Harden G, Ware D, Begum S, Mortimore B, Kerry G, Heath P, Phillimore B, Tracey A, Corby N, Dunn M, Johnson C, Wood J, Clark S, Pelan S, Griffiths G, Smith M, Glithero R, Howden P, Barker N, Lloyd C, Stevens C, Harley J, Holt K, Panagiotidis G, Lovell J, Beasley H, Henderson C, Gordon D, Auger K, Wright D, Collins J, Raisen C, Dyer L, Leung K, Robertson L, Ambridge K, Leongamornlert D, McGuire S, Gilderthorp R, Griffiths C, Manthravadi D, Nichol S, Barker G, Whitehead S, Kay M, Brown J, Murnane C, Gray E, Humphries M, Sycamore N, Barker D, Saunders D, Wallis J, Babbage A, Hammond S, Mashreghi-Mohammadi M, Barr L, Martin S, Wray P, Ellington A, Matthews N, Ellwood M, Woodmansey R, Clark G, Cooper JD, Tromans A, Grafham D, Skuce C, Pandian R, Andrews R, Harrison E, Kimberley A, Garnett J, Fosker N, Hall R, Garner P, Kelly D, Bird C, Palmer S, Gehring I, Berger A, Dooley CM, Ersan-Ürün Z, Eser C, Geiger H, Geisler M, Karotki L, Kirn A, Konantz J, Konantz M, Oberländer M, Rudolph-Geiger S, Teucke M, Lanz C, Raddatz G, Osoegawa K, Zhu B, Rapp A, Widaa S, Langford C, Yang F, Schuster SC, Carter NP, Harrow J, Ning Z, Herrero J, Searle SMJ, Enright A, Geisler R, Plasterk RHA, Lee C, Westerfield M, de Jong PJ, Zon LI, Postlethwait JH, Nüsslein-Volhard C, Hubbard TJP, Roest Crollius H, Rogers J, Stemple DL. 2013. The Zebrafish reference genome sequence and its relationship to the human genome. Nature 496:498–503. https://doi.org/10.1038/nature12111
11. Gomes MC, Mostowy S. 2020. The case for modeling human infection in Zebrafish. Trends Microbiol 28:10–18. https://doi.org/10.1016/j.tim.2019.08.005
12. Mostowy S, Boucontet L, Mazon Moya MJ, Sirianni A, Boudinot P, Hollinshead M, Cossart P, Herbomel P, Levraud J-P, Colucci-Guyon E. 2013. The Zebrafish as a new model for the in vivo study of *Shigella flexneri* interaction with phagocytes and bacterial autophagy. PLoS Pathog 9:e1003588. https://doi.org/10.1371/journal.ppat.1003588
13. Torraca V, Kaforou M, Watson J, Duggan GM, Guerrero-Gutierrez H, Krokowski S, Hollinshead M, Clarke TB, Mostowy RJ, Tomlinson GS, Sancho-Shimizu V, Clements A, Mostowy S. 2019. *Shigella sonnei* infection of Zebrafish reveals that O-antigen mediates neutrophil tolerance and dysentery incidence. PLoS Pathog 15:e1008006. https://doi.org/10.1371/journal.ppat.1008006
14. Willis AR, Torraca V, Gomes MC, Shelley J, Mazon-Moya M, Filloux A, Lo Celso C, Mostowy S. 2018. *Shigella*-induced emergency granulopoiesis protects Zebrafish larvae from secondary infection. mBio 9:e00933-18. https://doi.org/10.1128/mBio.00933-18
15. Van Ngo H, Robertin S, Brokatzky D, Bielecka MK, Lobato-Márquez D, Torraca V, Mostowy S. 2022. Septins promote caspase activity and coordinate mitochondrial apoptosis. Cytoskeleton. https://doi.org/10.1002/cm.21696
16. Zhou Z, Alikhan N-F, Mohamed K, Fan Y, Achtman M. 2020. The Enterobase user's guide, with case studies on salmonella transmissions, Yersinia Pestis Phylogeny, and Escherichia core Genomic diversity. Genome Res 30:138–152. https://doi.org/10.1101/gr.251678.119
17. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res 43:e15. https://doi.org/10.1093/nar/gku1196
18. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Res 46:e134. https://doi.org/10.1093/nar/gky783
19. Didelot X, Parkhill J. 2022. A scalable analytical approach from bacterial genomes to epidemiology. Philos Trans R Soc Lond B Biol Sci 377:20210246. https://doi.org/10.1098/rstb.2021.0246
20. Zhang X, Payne M, Nguyen T, Kaur S, Lan R. 2021. Cluster-specific gene markers enhance Shigella and enteroinvasive *Escherichia coli* in silico

serotyping. Microb Genom 7:000704. https://doi.org/10.1099/mgen.0.000704
21. Michelacci V, Tozzoli R, Arancia S, D'Angelo A, Boni A, Knijn A, Prosseda G, Greig DR, Jenkins C, Camou T, Sirok A, Navarro A, Schelotto F, Varela G, Morabito S. 2020. Tracing back the evolutionary route of enteroinvasive *Escherichia coli* (EIEC) and *Shigella* through the example of the highly pathogenic o96: H19 eiec clone. Front Cell Infect Microbiol 10:260. https://doi.org/10.3389/fcimb.2020.00260
22. Cowley LA, Oresegun DR, Chattaway MA, Dallman TJ, Jenkins C. 2018. Phylogenetic comparison of Enteroinvasive *Escherichia Coli* isolated from cases of Diarrhoeal disease in England, 2005-2016. J Med Microbiol 67:884–888. https://doi.org/10.1099/jmm.0.000739
23. Duggan GM, Mostowy S. 2018. Use of Zebrafish to study *Shigella* infection. Dis Model Mech 11: https://doi.org/10.1242/dmm.032151.
24. Lan R, Alles MC, Donohoe K, Martinez MB, Reeves PR. 2004. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* Spp. IAI 72:5080–5088. https://doi.org/10.1128/IAI.72.9.5080-5088.2004
25. Pilla G, McVicker G, Tang CM. 2017. Genetic plasticity of the *Shigella* virulence plasmid is mediated by Intra- and inter-molecular events between insertion sequences. PLoS Genet 13:e1007014. https://doi.org/10.1371/journal.pgen.1007014
26. Pasqua M, Michelacci V, Di Martino ML, Tozzoli R, Grossi M, Colonna B, Morabito S, Prosseda G. 2017. The intriguing evolutionary journey of Enteroinvasive *E. coli* (EIEC) toward pathogenicity. Front Microbiol 8:2390. https://doi.org/10.3389/fmicb.2017.02390
27. Sansonetti PJ, Kopecko DJ, Formal SB. 1982. Involvement of a plasmid in the invasive ability of *Shigella Flexneri*. Infect Immun 35:852–860. https://doi.org/10.1128/iai.35.3.852-860.1982
28. Qadri F, Hossain SA, Ciznár I, Haider K, Ljungh A, Wadstrom T, Sack DA. 1988. Congo red binding and salt aggregation as indicators of virulence in *Shigella species*. J Clin Microbiol. 26:1343–1348. https://doi.org/10.1128/jcm.26.7.1343-1348.1988
29. Connor TR, Loman NJ, Thompson S, Smith A, Southgate J, Poplawski R, Bull MJ, Richardson E, Ismail M, Thompson SE, Kitchen C, Guest M, Bakke M, Sheppard SK, Pallen MJ. 2016. CLIMB (the Cloud Infrastructure for Microbial Bioinformatics): an online resource for the medical microbiology community. Microb Genom 2:e000086. https://doi.org/10.1099/mgen.0.000086
30. Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313. https://doi.org/10.1093/bioinformatics/btu033
31. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. Nucleic Acids Res 46:e134. https://doi.org/10.1093/nar/gky783
32. Duchêne S, Duchêne D, Holmes EC, Ho SYW. 2015. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. Mol Biol Evol 32:1895–1906. https://doi.org/10.1093/molbev/msv056
33. Reinhardt J, Kolbe M. 2014. Secretion assay in *Shigella flexneri*. Bio-protocol 4. https://doi.org/10.21769/BioProtoc.1302

3.2.1. Manuscript supplementary material

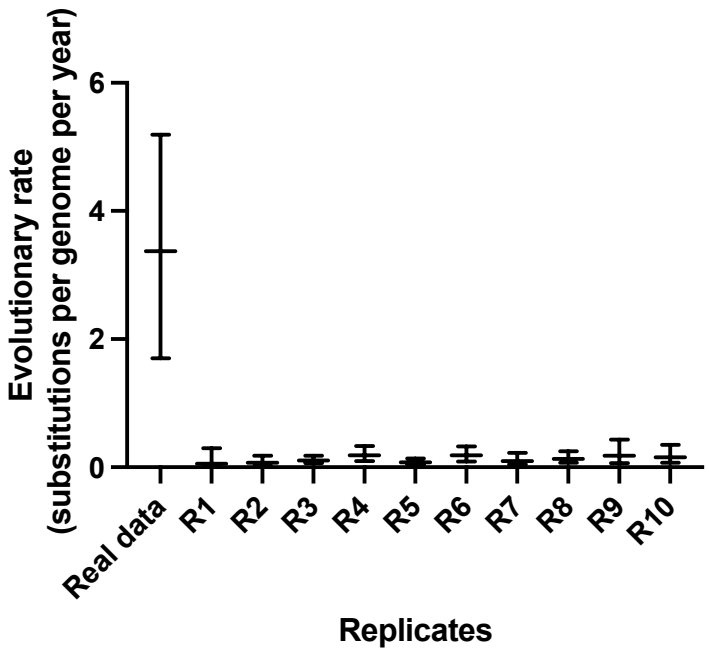


Figure S1. Results of date randomisation test. To test for temporal signal within our dataset, isolation dates were randomised n= 10 times and analysis was re-run (R1-R10). We saw no overlaps with the evolutionary rate of our real data, indicating the suitability of the dataset for performing a dated phylogenomic analysis.

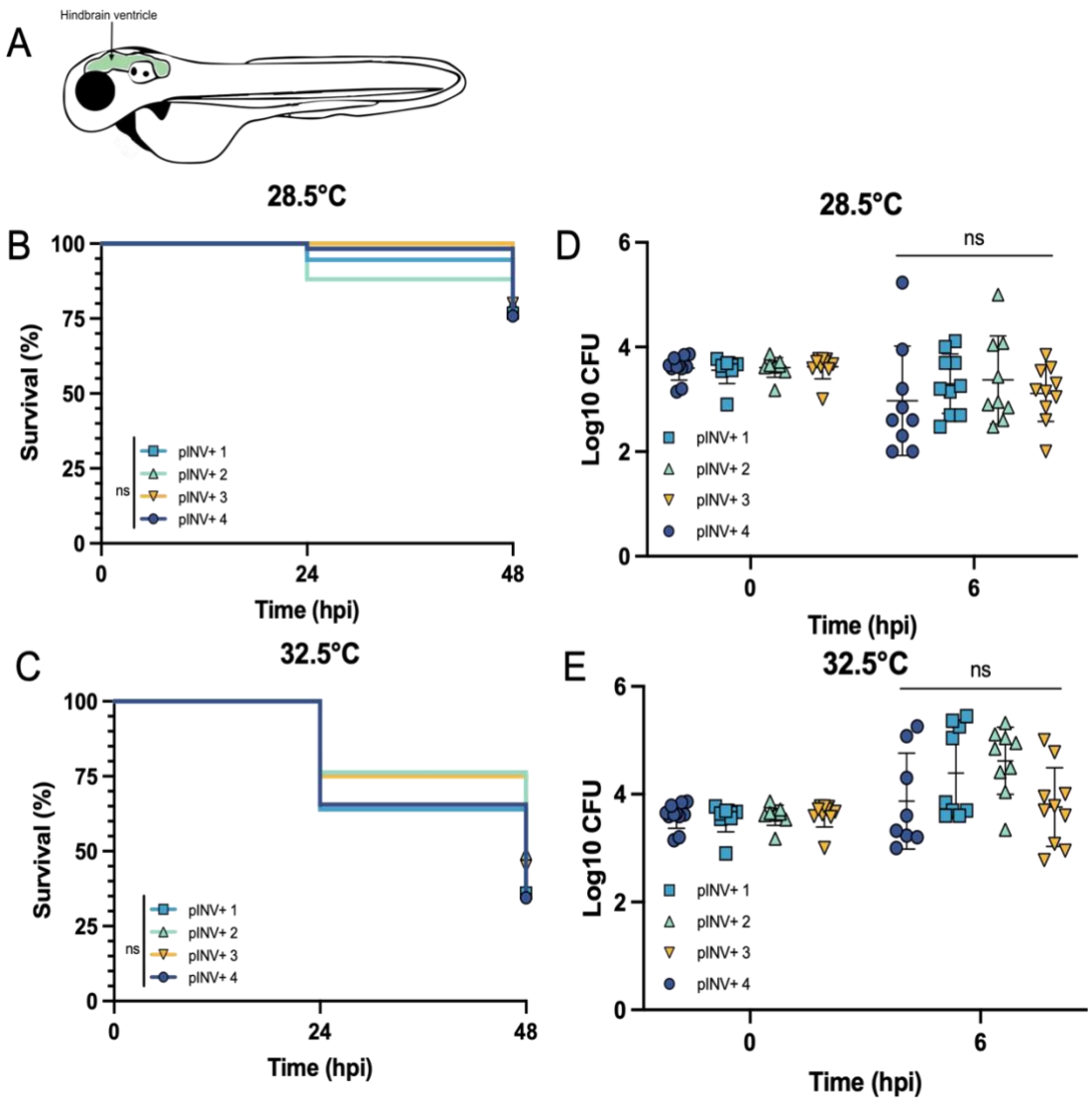


Figure S2. Virulence of ST99 EIEC isolates is temperature dependent. A) Schematic of a zebrafish larva, indicating the hindbrain ventricle as the infection site. **B-E)** 3-day post fertilisation larvae were injected with 5000 CFU of bacteria before being separated for incubation at 28.5 °C or 32.5 °C. **B,C)** pINV+ ST99 strains exhibit a temperature dependent virulence, with significantly more killing observed at 32.5 °C. No differences in survival were observed between strains (p= 0.22). Significance was tested using Log rank (Mantel-Cox) test. **D,E).** Enumeration of bacterial burden is also temperature dependent in pINV+ strains, with greater CFUs quantified at 6 hours post infection from larvae incubated at 32.5 °C. Significance for CFUs was tested using a two-way ANOVA with Sidak's correction.

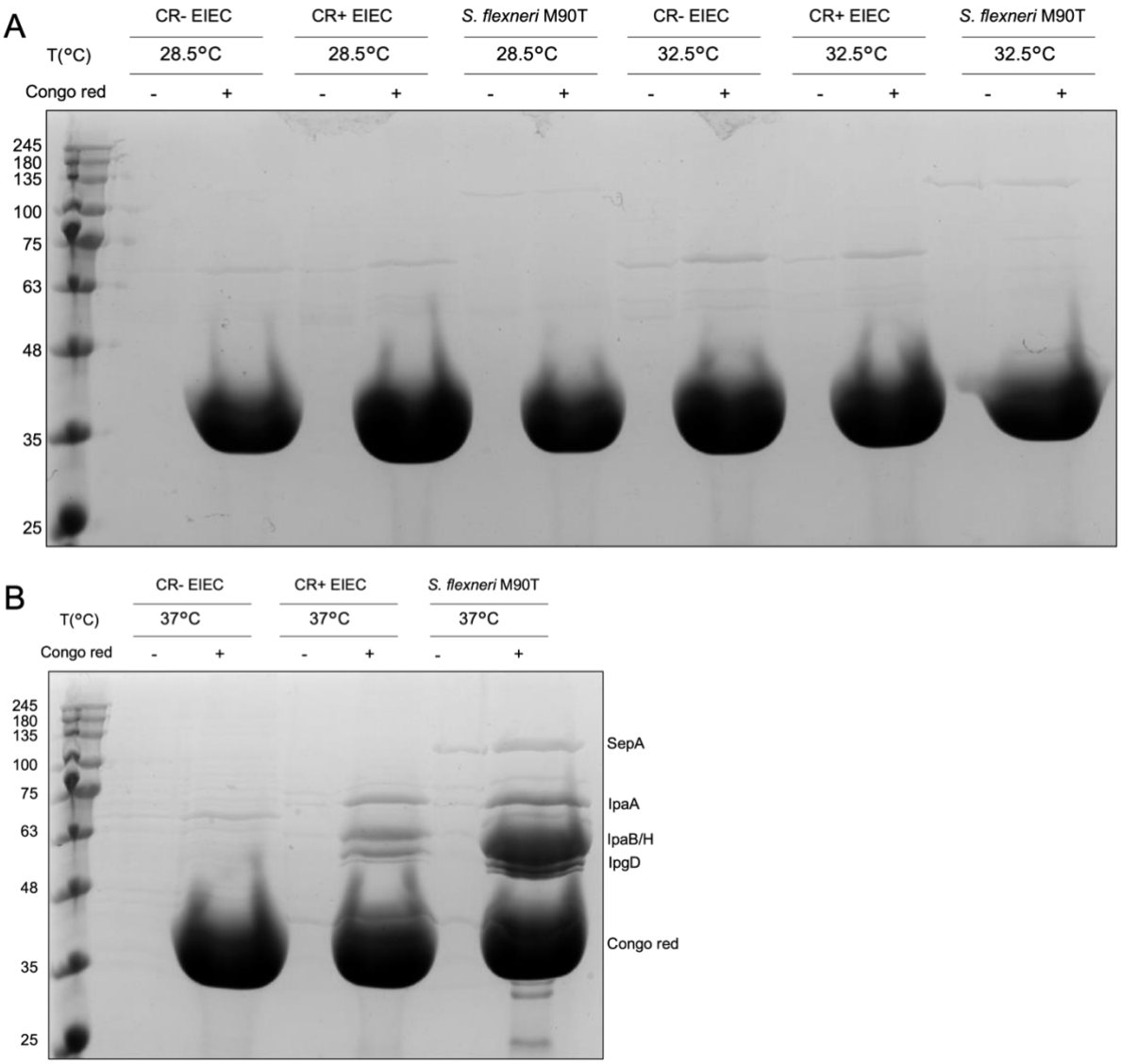


Figure S3. Secretion of virulence factors in vitro by ST99 EIEC pINV+1 (Congo red + colony), T3SS-deficient ST99 EIEC (Congo red - colony) and *S. flexneri*. SDS-PAGE gel stained with Coomassie blue, showing secreted factors in the presence or absence of Congo red, at different temperatures (**A**: 28.5 °C and 32.5 °C, **B**: 37 °C). *S. flexneri* M90T is used as a positive control. Several well characterised secreted virulence factors are identified and labelled based on their abundance and molecular weight.

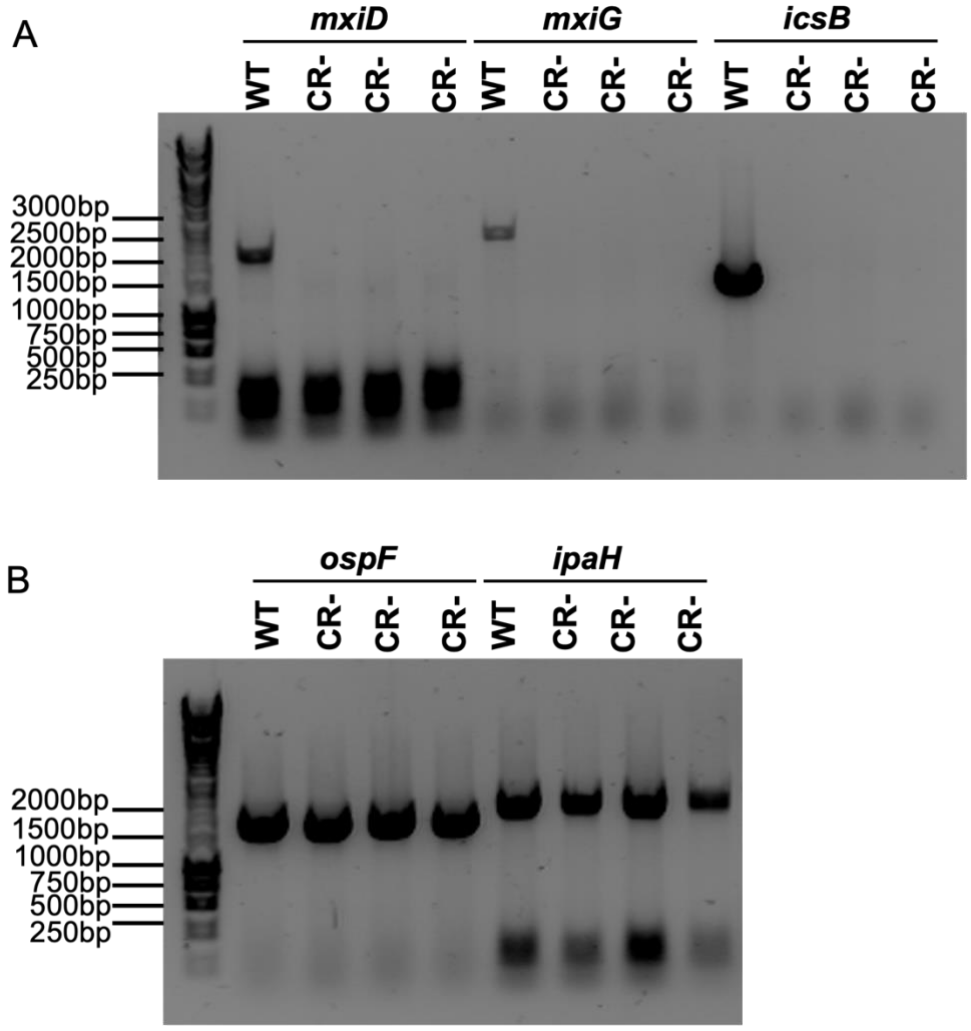


Figure S4. Colony PCR to check for the presence of pINV encoded genes. We performed colony PCR to check Congo red negative (CR-) pINV+1 colonies for **A)** the presence of genes located in the T3SS-encoding region (*mxiD*, *mxiG* and *icsB*) and **B)** outside of the T3SS-encoding region (*ospF* and *ipaH*). We see the loss of genes located in the T3SS-encoding region in CR- colonies, but not the wild type (WT), indicating that this region has been lost.

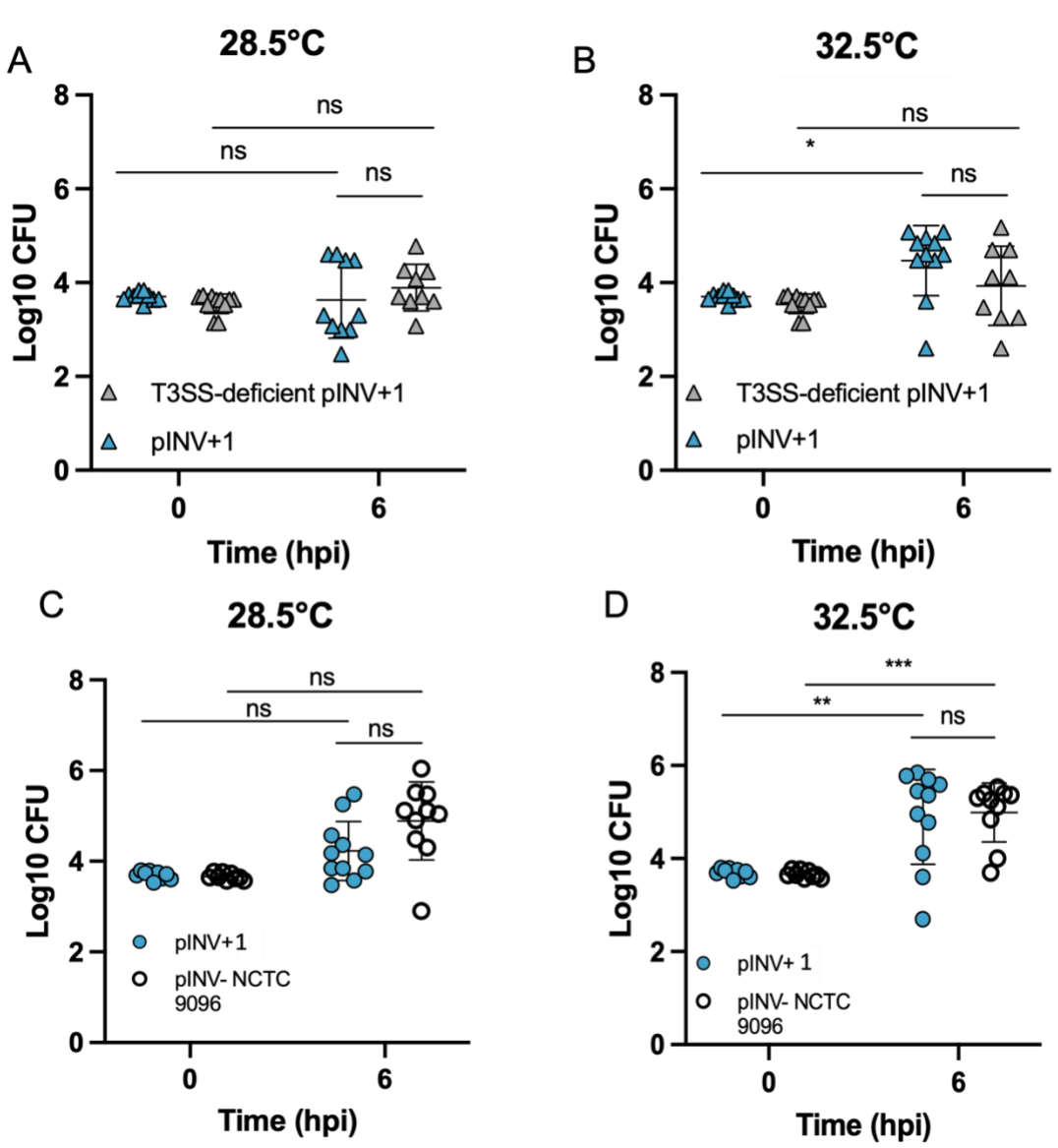


Figure S5. Bacterial burden of pINV+1 compared to its T3SS-deficient counterpart and the oldest available pINV- isolate, NCTC 9096. A,B) We observe a temperature dependent increase in CFUs enumerated from larvae infected with pINV+1 at 32.5 °C, but not for its T3SS-deficient counterpart. **C,D)** Enumeration of CFUs from infected larvae at 6 hpi is temperature dependent in the case of pINV+1, but not pINV- NCTC 9096. Significance is tested using a one-way ANOVA with Sidak's correction. *p<0.0332; **p< 0.0021; ***p<0.0002; ****p<0.0001.

Table S1 is too large to be inserted here and can be accessed at the following link: https://journals.asm.org/doi/suppl/10.1128/mbio.00882-23/suppl_file/mbio.00882-23-s0006.xlsx.

3.2.2. Supplementary results

Source of ST99 genomes used here

At the time of analysis, 92 ST99 whole genomes were publicly available on Enterobase. Of these, most were isolated from countries in the Western world (and considered to be HIC) (Fig 3.2.1A). The geographic distribution of isolates is likely more reflective of countries with an adequate pathogen genomic surveillance system in place rather than the actual epidemiology of ST99 *E. coli*.

As is true for *Shigella,* there is no known environmental or animal reservoir for EIEC, to test if this was true for the whole ST99 clone, the distribution of isolation sources for each genome used was examined. Most samples (60.87%) were found to be isolated from humans (likely to be from infection), but interestingly 30.21% of ST99 samples (all pINV negative and non-EIEC) were isolated from an animal or environmental source (Fig 3.2.1B). This suggests that prior to the acquisition of pINV, ST99 *E. coli* may have had a non-human niche, an implication which could be important for identifying the source of pINV.

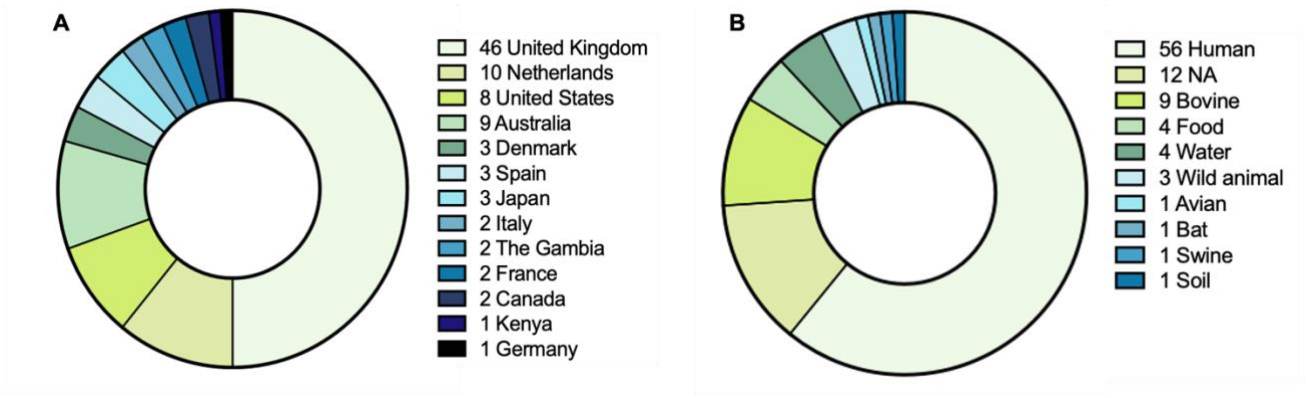


Figure 3.1. Distribution of location and source from which ST99 samples used for phylogenomic analysis were isolated. A) The location where samples were isolated from. **B)** The source from which ST99 samples were isolated. Each chart represents 92 samples in total.

Verification of temporal signal and dating model selection

To verify the presence of temporal signal within the ST99 dataset used here, a root-to-tip regression analysis was performed, this tests for an association between the number of nucleotide substitutions from the root and the isolation times of each sample. This analysis revealed a positive correlation between the evolutionary rate and time (R^2 = 0.19, Fig 3.2.2) which was found to be statistically significant (p = 0.006). This suggests a moderate temporal signal and thus indicates the suitability of the dataset for dating analysis.

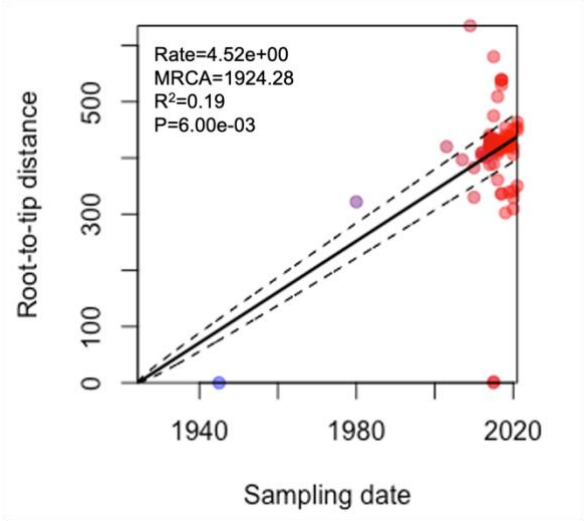


Figure 3.2. Root to tip regression analysis performed by BactDating. The root to tip analysis provides a rough estimate of temporal signal within the dataset. Here, we observe a positive correlation between genetic distance and time (R^2 value), which is supported statistically (p value). This indicates the presence of a moderate molecular clock signal within the dataset. The evolutionary rate is given in substitutions per genome per year.

To determine a suitable model for dating analysis, the performance of both strict and relaxed clock models on our dataset was evaluated. A strict model assumes that the same rate of evolutionary change is occurring in every branch, whereas a relaxed model allows each branch to have an independent evolutionary rate. The trace plots from each model (conducted by sampling every 100 iterations) showed convergence and mixing for both models (Fig 3.2.4), but tighter convergence can be observed for the relaxed clock model. To verify this observation, the 'modelcompare' function in BactDating, which compares models based on the deviance information criteria (DIC), was used. The DIC values were 2042.19 for the

relaxed model and 2709.27 for the strict model, confirming that the relaxed model is the most appropriate model to use for the dating analysis, in line with what is true for most other bacterial datasets (Duchêne et al., 2016).

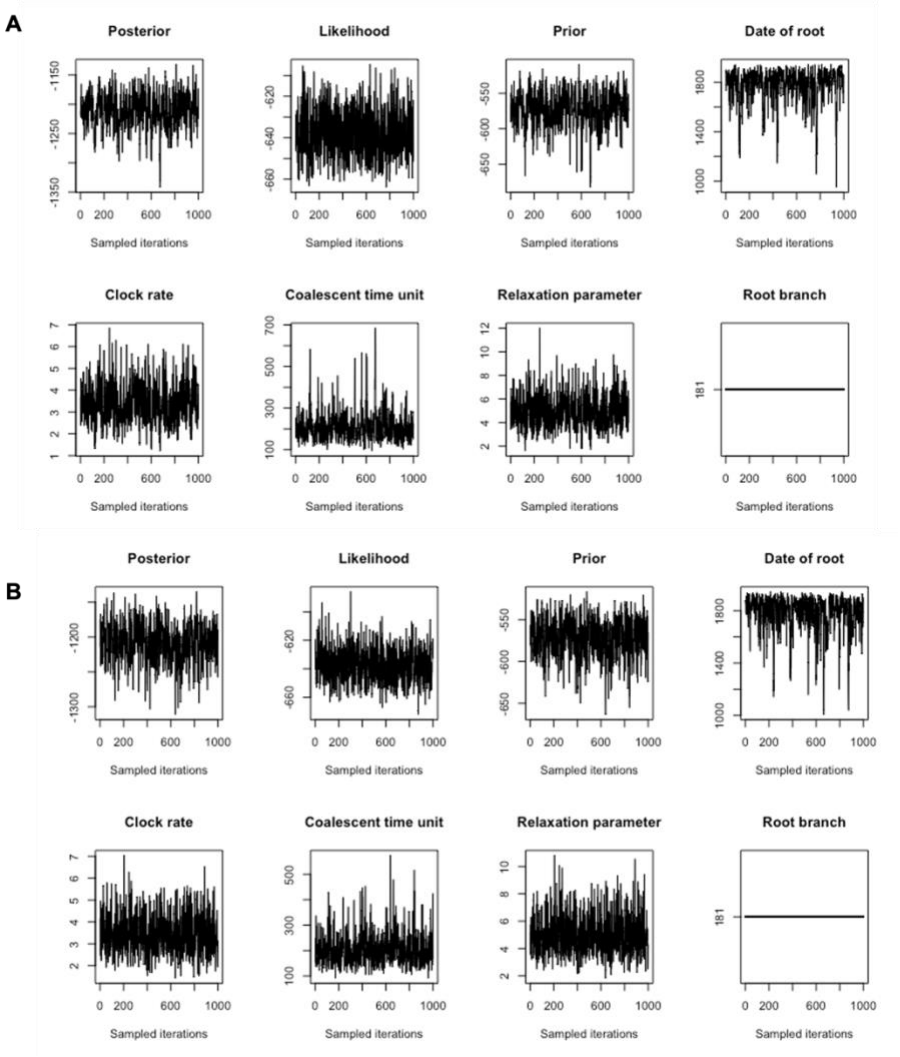


Figure 3.3. BactDating trace plots from the relaxed and strict clock models. Trace plots were conducted by sampling every 100 iterations over the MCMC run (with the first half excluded as burn in). A) Trace plot from 'relaxedgamma' model. B) Trace plot from 'strictgamma' model. Good mixing and convergence were observed for both models, but tighter convergence can be observed for the relaxed model and the results also passed the effective sample size test (with all values over 100). Clock rates are given in substitutions per genome per year.

Once the dating analysis had been performed with the previously identified relaxed model, the date randomisation test (results presented in section 3.2.1) was performed to verify the presence of temporal signal and to ensure that the results obtained were significant. This test involves randomly reassigning dates of isolation and re-running the analysis at least 10 times, if the results overlap with the real dataset, then the temporal signal within the dataset cannot be considered significant. An evolutionary rate of 4.52 substitutions per genome per year (equivalent to 8.92×10^{-7} substitutions per site per year) was estimated for the real dataset. This value is consistent with the evolutionary rate of the *S. sonnei* dataset used in the validation of BactDating, which was found to have an evolutionary rate of 4.22 substitutions per genome per year (Didelot et al., 2018), and with previous studies that have reported a substitution rate of 6×10^{-7} per site per year for *S. sonnei* (Holt et al., 2012) , 8.7×10^{-7} for *S. dysenteriae* (Njamkepo et al,. 2016) and 9.5×10^{-7} *for S. flexneri* (Connor et al., 2015).

We saw no overlap in the substitution rate of real and randomised data (Fig S1) and as an additional measure, saw that there was no mixing or convergence of the MCMC traces when randomised datasets were used (Fig 3.2.4). Finally, the comparison of DIC values also agreed with these findings, with a value of 4040.20 obtained for the randomised dataset (as compared to 2042.19 for the real dataset), showing that the results are significantly different.

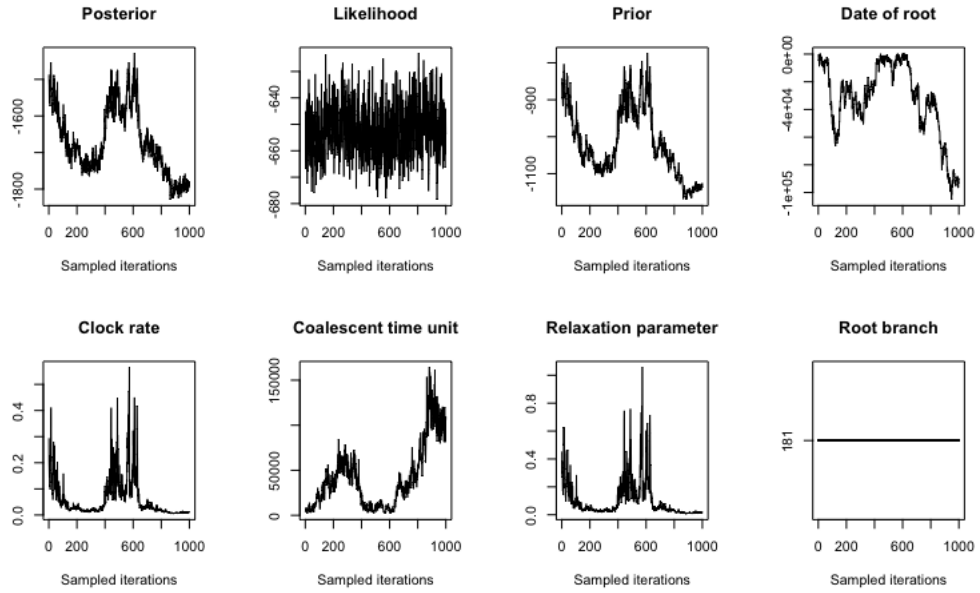


Figure 3.4. BactDating trace plots from the randomised dates dataset with a relaxed model. Trace plots were conducted by sampling every 100 iterations over the MCMC run (with the first half excluded as burn in). For the randomised dates test, no mixing and convergence was observed for most parameters, as expected. Clock rates are given in substitutions per genome per year.

3.3. Discussion

3.3.1. Overview

The recent emergence of a novel EIEC clone provided a unique opportunity to understand the early stages of EIEC and *Shigella* evolution. In this chapter, phylogenomic analysis was used to explore the evolution of ST99 *E. coli* and estimate a date for the emergence of the EIEC clone. Zebrafish infection has previously been valuable in revealing differences in virulence between different *Shigella* species (Torraca et al., 2023, Torraca et al., 2019). Here, zebrafish infection was used to complement bioinformatic methodologies, revealing distinct mechanisms of virulence between phylogenomic clusters.

3.3.2. ST99 EIEC diverged ~40 years ago

Phylogenomic analysis of all publicly available ST99 genomes (inclusive of EIEC and non-EIEC isolates) revealed that those isolates harbouring pINV, and thus considered to be EIEC, formed a distinct cluster compared to non-EIEC ST99 genomes. The presence of a single defined cluster indicates that pINV has been acquired on one occasion by a single common ancestor, concomitant with monophyletic *Shigella* subgroups, such as *S. sonnei* (Holt et al., 2012). Within the EIEC cluster, there are several tight clusters (with short branch lengths) which may be representative of regional dispersal, or strains isolated from different outbreaks. Generation of a time-calibrated phylogeny illuminated the relatively short evolutionary history of ST99 EIEC, with a possible most recent common ancestor (MRCA) existing ~1982. This date is in line with suggestions of a recent acquisition of pINV (compared to other *Shigella* and EIEC subgroups) and explains previous findings highlighting a lack of adaptive mutations and metabolic phenotypes resembling commensal *E. coli* (Michelacci et al., 2016).

3.3.3. ST99 EIEC virulence is temperature-dependent in zebrafish

The thermoregulation of virulence at the transcriptional level is well documented in *Shigella,* and other host-adapted pathogens (Falconi et al., 1998, Waldminghaus et al., 2007). In this chapter, the thermoregulated virulence of a novel EIEC clone is demonstrated for the first time *in vitro* and *in vivo* and is attributed solely to the presence of a functional T3SS. This indicates

that despite ST99 EIEC representing a relatively early stage of evolution (and certainly the earliest documented to date), the coordination of virulence genes in response to a host environment is still tightly regulated, marking pINV acquisition as a pivotal step in this process. Additionally, the T3SS activity (measured by the abundance of secreted effector proteins) was found to be lower than *S. flexneri* overall, consistent with the hypothesis that EIEC causes less severe infections (van den Beld and Reubsaet, 2012) and represents an evolutionary intermediate.

3.3.4. ST99 *E. coli* utilise temperature-dependent and -independent mechanisms of virulence

The *Shigella*-zebrafish model has been used previously to reveal differences in virulence between *S. flexneri* and *S. sonnei* (Torraca et al., 2019), data presented here further explores differences in virulence between isolates from the same ST and of the same serotype, highlighting the power of the model to resolve such fine scale differences. It was found that an older, non-EIEC isolate did not exhibit thermoregulated virulence (as expected), but surprisingly was just as virulent as an EIEC isolate through a T3SS-independent mechanism yet to be elucidated. Previously it has been assumed that the transition from *E. coli* to *Shigella* or EIEC begins with an innocuous *E. coli* ancestor, however, our findings contradict this and instead suggest that an already pathogenic *E. coli* acquired pINV. In future, a finer scale analysis of non-EIEC ST99 isolates would be of interest to identify and explore factors that might pre-dispose a non-invasive *E. coli* to acquiring pINV and beginning the process of host specialisation.

3.3.5. Conclusions

In conclusion, this study demonstrates the short evolutionary history of ST99 EIEC, confirming a recent acquisition of pINV on a single occasion. A zebrafish infection model was then used to explore the virulence of EIEC and non-EIEC ST99 isolates, illuminating two independent mechanisms of virulence and verifying an important role for pINV acquisition in the early stages of EIEC and *Shigella* emergence.

Chapter 4. Comparative genomic analysis highlights evidence of ongoing adaptation in lineage 3 *Shigella sonnei*

4.1. Introduction

4.1.1. General features of *S. sonnei*

S. sonnei can be differentiated from the other *Shigella* subgroups through its numerous atypical properties, most notably its surface polysaccharides. LPS is an important component of the Gram-negative bacterial cell envelope and a key factor in bacterial virulence. In *S. sonnei,* it has been implicated in bacterial resistance of phagocytic killing and hindering host cell death for its own survival advantage (Watson et al., 2019, Torraca et al., 2019). LPS is comprised of a lipid core (which attaches to the outer membrane), a core oligosaccharide (which displays limited variability) and the O-antigen (which contains polysaccharide repeats and can be highly variable within and between species) (Erridge et al., 2002). In *Shigella* and *E. coli*, the production of LPS is typically controlled by the Wzx/Wzy dependent pathway which synthesises, assembles and exports polysaccharide. The genes involved in LPS synthesis are encoded on the chromosome for all *Shigella* subgroups, except for *S. sonnei*, where it is instead encoded on pINV. The structure of *S. sonnei* LPS is also uncommon, with its O-antigen chain comprised of repeating *N*-acetyl-$_{L}$-altrosaminuronic acid ($_{L}$-AltNAcA) and 4-amino-4-deoxy-*N*-acetyl-$_{D}$-fucosamine (FucNAc4N) residues, sugars which are not found elsewhere in the *E. coli* or *Shigella* population. A detailed analysis of the *S. sonnei* O-antigen gene cluster revealed that it is almost identical to the O-antigen present in serotype O17 *Plesiomonas shigelloides*, with phylogenetic analysis highlighting a recent acquisition of these genes by *S. sonnei*, likely through horizontal gene transfer (Shepherd et al., 2000). This study also investigated the chromosomal *wzy* locus, finding that it is present but disrupted within *S. sonnei*, suggesting that *S. sonnei* acquired its new O-antigen gene cluster via horizontal gene transfer, which was then followed by the disruption of the chromosomal locus, perhaps driven by a differential niche adaptation.

In addition to a unique LPS, an O-antigen capsule has also been described for *S. sonnei*, making it the only capsulated *Shigella* subgroup described to date. The capsule belongs to the 'group 4' capsule (G4C) family, and it is structurally similar to the O-antigen (Caboni et al., 2015) (Fig 4.1), comprising of a lipid anchor and polysaccharide repeats identical to those present in the LPS O-antigen. In the case of *S. sonnei,* the O polysaccharide synthesis relies on the pINV encoded O-antigen synthesis cluster, but the capsule transport machinery is chromosomally encoded. The capsule was found to confer resistance to complement mediated killing and bacterial dissemination in a rabbit infection model, but these advantages came at the expense of invasiveness. The G4C was also found to physically mask surface structures, such as the T3SS, and subsequently, a capsule mutant was shown to be significantly more invasive in tissue culture cells (Caboni et al., 2015).

Both the LPS and the G4C rely on pINV encoded O-antigen synthesis machinery, however, pINV in *S. sonnei* is highly unstable in culture, resulting in an avirulent phenotype when lost (Sansonetti et al., 1981). The increased frequency of pINV loss in *S. sonnei* (compared to other *Shigella* subgroups) has been attributed to the IS mediated disruption of toxin-antitoxin systems (VapBC, GmvAT and CcdAB) that aid in plasmid maintenance through post-segregational killing (Martyn et al., 2022). The instability of *S. sonnei* pINV has thus hampered the study of *S. sonnei* virulence, with a significant proportion of available *S. sonnei* genomes lacking pINV sequences (Holt et al., 2012) and a stark lack of work using infection models as compared to *S. flexneri*, which loses pINV at a much lower frequency.

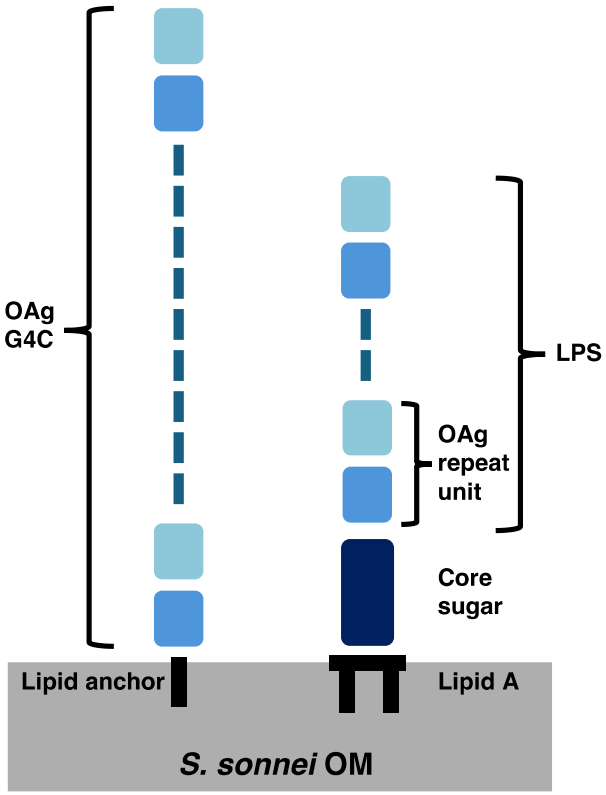


Figure 4.1. Schematic of the surface polysaccharides present on the outer membrane (OM) of *S. sonnei*. Lipopolysaccharide (LPS) is comprised of a Lipid A anchor, a core polysaccharide and O-antigen (OAg) chain repeats (alternating repeats of $_{L}$-AltNAcA and FucNAc4N). The O-antigen chain length in *S. sonnei* is thought to be monomodal, comprised of 20-25 repeating units. *S. sonnei* also has a group 4 capsule (G4C), also known as an O-antigen capsule, which is comprised of a lipid anchor and polysaccharide repeats identical to those present in the LPS O-antigen. The G4C is of a high molecular weight, comprised of over 100 repeating polysaccharide units.

4.1.2. *Shigella sonnei* population structure

S. sonnei is the least genetically diverse of the *Shigella* subgroups, comprising just a single serotype with a highly clonal population structure (Muthuirulandi Sethuvel et al., 2017). *S. sonnei* is delineated into five main lineages (1-5, (Fig 4.2)), with a single common ancestor predicted to have existed ~500 years ago (Holt et al., 2012). A hierarchal genotyping scheme based on single nucleotide variations (SNV) has been developed for *S. sonnei*, simplifying the identification and nomenclature of lineages and facilitating comparison between studies (Hawkey et al., 2021). The *S. sonnei* genotyping scheme, implemented in the Mykrobe software, separates each main lineage by ~600 SNVs, clades by ~215 SNVs and subclades by ~100 SNVs. Among the five lineages, there have been varying degrees of global dissemination and expansion: lineage 1 is considered the most ancestral lineage and is infrequently detected outside of Europe; lineage 4 represents an extinct lineage, comprising of just a single known isolate (Holt et al., 2012); and lineage 5 is restricted to Latin America and parts of Africa (Baker et al., 2017). Lineage 2 has undergone limited dissemination, establishing localised clones in some regions of Africa, Asia and South America, but overall, lineage 3 has been the most successful at global dissemination having been detected on every continent, and today represents the most epidemiologically significant lineage (Hawkey et al., 2021).

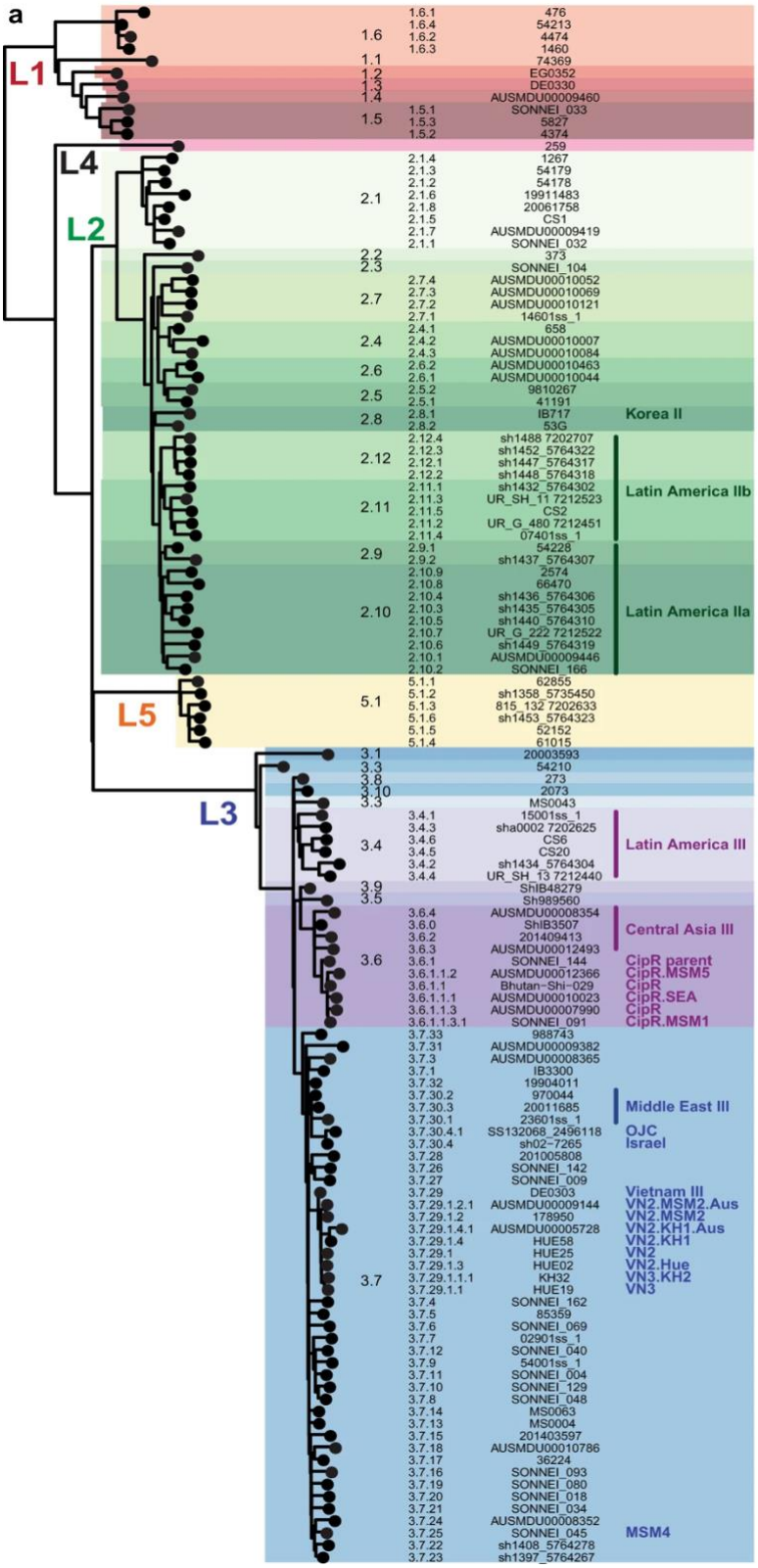


Figure 4.2. Maximum likelihood phylogenetic tree of representative *S. sonnei* genomes. The *S. sonnei* population can be delineated into 5 distinct lineages (L1-5) separated by ~600 single nucleotide variants, which can be further split into clades and subclades. Epidemiologically important clades are denoted in bold in the far-right column. Figure obtained from Hawkey et al, 2021.

The increase in lineage 3 *S. sonnei* was first observed in the 1980s and regional studies into its evolution unveiled a pattern of clonal replacement and expansion once introduced into a new region, seemingly driven by the independent acquisitions of genes conferring multidrug resistance (encoded on a class II integron (Int*2*)-bearing transposon, Tn7) (Holt et al., 2013, Chung The et al., 2021). Within lineage 3, clades 3.6 (also known as the Central Asia/CipR clade) and 3.7 (known as Global 3) dominate the epidemiological landscape, accounting for most contemporary infections (Hawkey et al., 2021). Clade 3.6 is associated with circulation in South and Central Asia and gave rise to several subclades that are resistant to fluoroquinolones (conferred by a triple mutation in the quinolone resistance determining regions (QRDR) of *parC* and *gyrA*), which have since disseminated globally (Chung The and Baker, 2018, Chung The et al., 2019). More recently, subclade 3.6.1.1.2 has been associated with the rapid spread of XDR *S. sonnei* among MSM, following the acquisition of resistance to 3rd generation cephalosporins, macrolides and β-lactams (Charles et al., 2022, Mason et al., 2023). The sustained transmission of clade 3.7 has similarly been observed globally, with subclades exhibiting tight regional associations (Hawkey et al., 2021). A study of Vietnamese *S. sonnei* infections highlighted a strong selection for the accumulation and fixation of both AMR determinants and a colicin system within subclade 3.7.29 (Holt et al., 2013). A similar pattern was observed for subclade 3.7.30.4, which has been linked to several large-scale outbreaks among the Orthodox Jewish communities in Europe and North America (Baker et al., 2016).

Increased resistance to antimicrobials is a clear signature of lineage 3 *S. sonnei*, but there is also evidence to suggest adaptive evolution is occurring elsewhere in the genome. There is limited gene content variation between *S. sonnei* lineages, but lineage 3 was found to have an increased nucleotide substitution rate (the number of nucleotide mutations that become fixed per site over time) as compared to the global *S. sonnei* phylogeny (Holt et al., 2012, Chung The et al., 2019). Additionally, lineage 3 displays an increased abundance of chromosomal IS elements (Hawkey et al., 2020), suggestive of ongoing evolutionary

adaptation, potentially driven by different selection pressures. A trend in the loss of metabolic phenotypes and immunogenic features has also been highlighted, with interruption of *lpfC* (part of a fimbrial operon) documented throughout lineage 3 (Holt et al., 2012). Additionally, within clade 3.6, there is evidence of increased resistance to oxidative stress, and selection towards SNP accumulation in genes associated with stress response (Chung The et al., 2019). Overall, there is some evidence of ongoing adaption within the genomes of lineage 3 isolates from previous studies, but this has not been investigated comprehensively.

4.1.3. The dynamic *Shigella* genome

A fundamental *Shigella* genome is comprised of a single circular chromosome and pINV, but advances in whole genome sequencing have illuminated the genome plasticity and diversity between *Shigella* subgroups and lineages, comprising a core genome (of ~2000 genes) and a more mosaic-like collection of accessory elements (Yang et al., 2005). Genome plasticity in *Shigella* is principally mediated by the large abundance of ISs (small sequences of DNA encoding a transposase, catalysing their own mobility) which can facilitate genome rearrangements and the modulation of gene expression (Schneider et al., 2000). The mobility and expansion of ISs can also result in the formation of pseudogenes (coding sequences which may be disrupted and do not code for a protein, despite resembling a functional gene); pseudogenisation has been implicated in the convergent disruption of anti-virulence genes in *Shigella* (Jin et al., 2002). IS mediated deletions often occur alongside translocations and inversions, and rearrangement events have been documented in *Shigella* genomes of all subgroups (Yang et al., 2005). Bacteriophage-mediated gene transfer has also been imperative in shaping the diversity of *Shigella*, resulting in the integration of the pathogenicity islands of *Shigella,* which enhance virulence and aid in bacterial fitness and survival (Parajuli et al., 2017). Aside from pINV, *Shigella* is also known to access the wider Enterobacteriaceae gene pool, acquiring many smaller plasmids that can encode for AMR determinants and colicin systems, allowing different subgroups/lineages to adapt to the potentially different niches they occupy (Locke et al., 2021, Malaka De Silva et al., 2022). Together, through the convergent

loss and gain of functions, the dynamic nature of the *Shigella* genome has facilitated its adaptation to the human host.

4.1.4. Bacterial genome sequencing

The rapid advance and increasing accessibility of next-generation sequencing have been instrumental in resolving the broader population structure and exploring the genetic diversity of *Shigella* (Baker et al., 2024). Short read sequencing, which is dominated by Illumina technology, has been most commonly used for prokaryotic genome sequencing due to its low cost per base and highly accurate read outs. Short read sequencing typically produces reads of up to 300 basepairs (bp), which can impede the assembly of bacterial genomes and the analysis of structural variation, especially in the case of *Shigella* where there is an abundance of repetitive IS elements which are often longer than 300 bp (Phillippy et al., 2008). Long read sequencing, commonly enabled by PacBio or Oxford Nanopore Technologies (ONT), can be used to overcome such limitations. Long read sequencing produces read lengths typically exceeding 10,000 bp which are long enough to cover repetitive elements and enable the assembly of bacterial genomes to completion if read depth is sufficient (Koren and Phillippy, 2015). Long reads can, however, underrepresent very small plasmids and are prone to a higher error rate than short reads, although the latter limitation is improving rapidly as sequencing chemistries and base calling software are updated (Wick et al., 2023).

De novo assembly, a method for reconstructing a genome based only on overlapping DNA reads, can incorporate the benefits of both short and long read sequencing to generate a complete and highly accurate bacterial genome. The most accurate method of bacterial genome assembly at present involves a long read first approach, whereby a first assembly with no major structural errors is obtained, which is then followed by polishing with highly accurate short reads to remove any small-scale errors (Wick et al., 2021). To date, there are currently over 19,000 publicly available genomes of *S. sonnei*, but only 65 have contiguous chromosomal sequences (Baker et al., 2024), with even fewer containing pINV sequences,

meaning structural and IS mediated diversity, as well as pINV variability, have been relatively unexplored for *S. sonnei*.

4.1.5. Aims

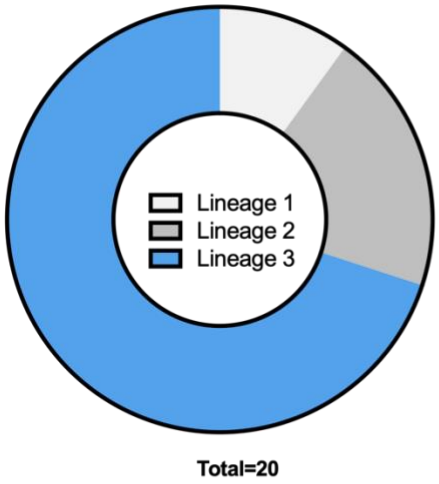
This chapter investigates genomic variations in a newly sequenced collection of epidemiologically relevant *S. sonnei* isolates, with a broad aim to identify any lineage-dependent variations that contribute towards the domination of lineage 3 in the *S. sonnei* epidemiological landscape. The specific aims of this chapter were: to collate a subset of epidemiologically representative *S. sonnei* isolates; to generate their complete, high quality reference genomes; and to compare completed genomes to identify lineage dependent differences in genome content and genome structure.

4.2. Results

4.2.1. Assembling a collection of representative *S. sonnei* isolates

First, a collection of *S. sonnei* isolates was identified to represent the diversity of the current *S. sonnei* population, based on existing short-read (Illumina) data. A target of 35 clinical isolates, including representatives from each major lineage and epidemiologically important sub-lineage, was initially set. Of these, a representative from lineage 4 was not attainable since there was only one extant strain available, which has since become non-viable during more than 50 years of laboratory storage (François Xavier-Weill, Institut Pasteur Enteric Bacterial Pathogens Unit, personal communication). Two different representatives from lineage 5 were obtained for culture, but multiple rounds of sub-culturing revealed that they were both pINV negative and thus not suitable for inclusion in the collection, with no alternative lineage 5 representatives in the accessible strain collections (which were already reduced due to the limitations of strain sharing during the COVID-19 pandemic). Of the remaining 32 isolates, pINV positive colonies were recovered for 20 isolates, which comprised two lineage 1, four lineage 2 and 14 lineage 3 isolates (Fig 4.3), covering most epidemiologically important variants of *S. sonnei*.

A



B

Genotype	Epidemiological importance
1	
1.5	
2.1	
2.3	
2.8/53G	Associated with Korea, lab-adapted
2.12	Associated with Latin America
3.4	Associated with Latin America
3.6.1	Ciprofloxacin resistance emerged
3.6.1.1.1	Triple QRDR mutation
3.6.1.1.3.1	MSM clade 1
3.6.2	Associated with Central Asia
3.7.3	
3.7.11	
3.7.16	
3.7.25	MSM clade 4
3.7.28	
3.7.29.1	Associated with Vietnam
3.7.30.1	Associated with the Middle East
3.7.30.4	Associated with Israel
3.7.30.4.1	Associated with OJC

Figure 4.3. Summary of *S. sonnei* genotypes included in the isolate collection for this study. A) A total of 20 virulence plasmid (pINV) positive clinical isolates were recovered for inclusion in the isolate collection, comprising of two lineage 1, four lineage 2 and 14 lineage 3 isolates. **B)** List of genotypes included in the isolate collection. The epidemiological importance column refers to epidemiological clusters (either by regional association, drug resistance phenotype, or circulation within the men who have sex with men (MSM) community), as summarised previously (Hawkey et al., 2021). QRDR = quinolone resistance determining region, OJC = Orthodox Jewish Community.

4.2.2. Generation of complete genomes for representative *S. sonnei* lineages

All isolates in the collection were subjected to both long-read and short-read sequencing (on the same, fresh DNA extracts) and a hybrid genome assembly approach was used to generate complete genome sequences (except for lab strain 53G (genotype 2.8) which already had a completed genome including pINV, generated by traditional Sanger sequencing, GenBank accession GCA_000283715.1) (Table 4.1). Of the 19 isolates that were newly sequenced, a circularised chromosome was achieved for 17 isolates (the two isolates for which a circularised

chromosome was not achieved belonged to genotypes 3.7.3 and 3.7.25). MLST analysis, (using the Achtman 7-locus scheme) was then performed on all completed genomes to verify sample identity. Eighteen isolates were positively identified as *S. sonnei* ST152 but one isolate (predicted to be genotype 3.6.1.1.3.1 from previous short read sequencing) was MLST typed as *S. boydii* ST145; this isolate was therefore excluded from further analyses. One of the ST152 genomes was not assigned a genotype by the Mykrobe *S. sonnei* genotyping scheme; this isolate fits within the *S. sonnei* phylogeny as a lineage 1 but lacks the *uid* operon, which Mykrobe uses to confirm *S. sonnei* status before subjecting reads to genotyping against the *S. sonnei* scheme (from personal communication with François-Xavier Weill, Institut Pasteur, owner of the isolate and Jane Hawkey, developer of the Mykrobe *S. sonnei* genotyping panel). This isolate was therefore considered a lineage 1 for the purposes of this analyses. Complete pINV sequences were present for 15 of the newly sequenced isolates, but pINV was found to be missing from two isolates (genotypes 3.7.25 and 3.7.30.4), likely due to plasmid loss during culture for sequencing which is frequent in *S. sonnei*. To quality check assemblies, completeness and contamination were assessed using CheckM. Completeness was > 97% for all assemblies (> 99% for most) and most genomes had < 1% contamination (apart from the two incomplete genomes, which had < 4%), indicating that they can be considered 'near complete' (the highest category of completeness) according to the guidelines given by CheckM (Parks et al., 2015).

Complete chromosome sizes were variable, ranging between 4.65 and 4.98 Megabase pairs (Mbp) and this finding was also reflected in the number of coding sequences (CDSs) annotated by Bakta, which ranged between 4747 and 5071 (Fig 4.4). The average chromosome sizes for each lineage were similar overall: lineage 1 had an average chromosome size of 4.90 Mbp; lineage 2 chromosomes averaged at 4.96 Mbp whilst lineage 3 had the smallest average chromosome size of 4.84 Mbp. The average number of CDSs annotated in each lineage echoed the chromosome sizes: with an average of 4935 predicted for lineage 1; 5025 for lineage 2 and 4931 for lineage 3.

pINV sizes ranged between 212 and 242 kilobase pairs (kbp) and some lineage-dependent variation in size was observed: lineage 1 isolates had a larger pINV size on average, at 240 kbp; lineage 2 had an average pINV size of 219 kbp; and lineage 3 had the smallest average pINV size at 214 kbp. Like with the chromosomal CDSs, this trend was also reflected in pINV CDSs: where an average of 272 were predicted for lineage 1; 242 for lineage 2 and 240 for lineage 3. Together, both chromosome and pINV sizes were smallest on average in lineage 3 genomes, in line with expectations from prior reports on genome reduction, although it is important to note that the differences reported here are small and verification in a larger dataset will be important in future.

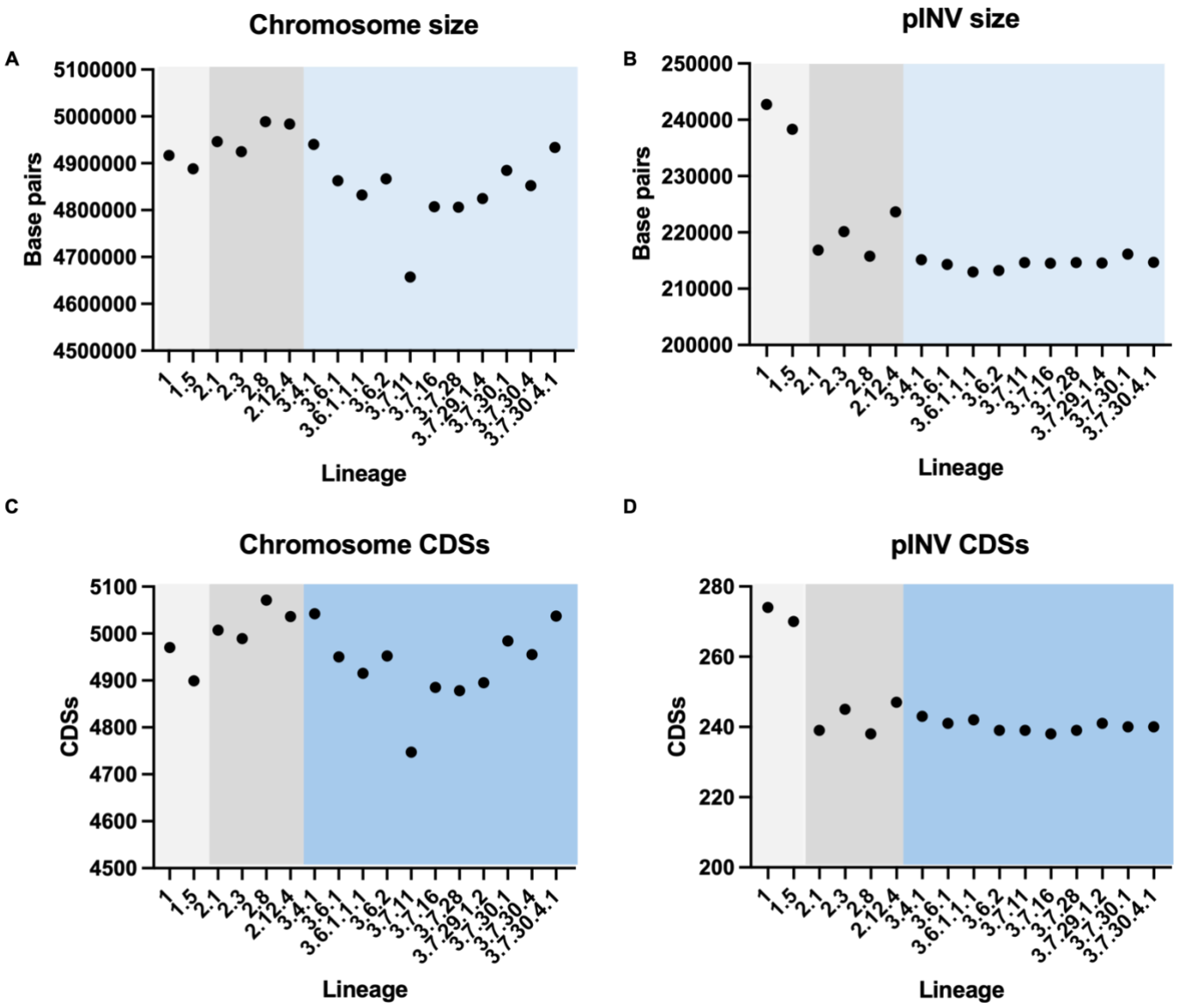


Figure 4.4. The total size and abundance of coding sequences (CDSs) in *S. sonnei* chromosomal and pINV sequences. A) Chromosome sizes of all *S. sonnei* assemblies with circularised chromosomes. **B)** pINV sizes of all *S. sonnei* assemblies with circularised pINV sequences. **C)** Number of CDSs annotated by Bakta for all *S. sonnei* assemblies with circularised chromosomes. **D)** Number of CDSs annotated by Bakta for all *S. sonnei* assemblies with circularised pINV sequences. CDSs were annotated using the full database option in Bakta. Lineage 1 isolates are highlighted in light grey, lineage 2 in dark grey and lineage 3 in blue.

Table 4.1. General features of *S. sonnei* genome assemblies presented here. [1]Genotypes were predicted using the *S. sonnei* genotyping framework implemented in Mykrobe, [2]CDSs,tRNAs and rRNAs were predicted using Bakta, [3]Completeness and contamination were assessed using CheckM. Features of the completed genome of lineage 2.8 are included for comparative purposes.

Genotype[1]	Genome size (bp)	N50	GC (%)	Depth	CDSs[2]	tRNAs	rRNAs	Contigs	Chromosome size (bp)	pINV size (bp)	Other plasmids	Complete-ness[3]	Contamina-tion[3]
Undetected (1)	5389660	4916714	50.8	61x	5428	97	22	7	4916714	242731	5	99.54	0.26
1.5	5260824	4888493	50.9	59x	5255	97	22	4	4888493	238313	2	99.54	0.16
2.1	5170588	4946399	50.8	76x	5176	97	22	4	4946399	216858	2	99.56	0.11
2.3	5150027	4924717	50.8	60x	5166	96	22	3	4924717	220157	1	98.98	0.11
2.8	5220473	4988504	50.7	NA	5248	96	22	5	4988504	215774	3	NA	NA
2.12.4	5289501	4983563	50.8	47x	5306	93	22	4	4983563	223662	2	99.6	0.11
3.4.1	5174681	4940164	50.8	74x	5229	97	22	7	4940164	215149	5	99.6	0.11
3.6.1	5112917	4862817	50.8	68x	5154	97	22	9	4862817	214316	7	99.6	0.26
3.6.1.1.1	5073758	4832017	50.8	86x	5110	96	22	5	4832017	212976	3	99.6	0.11
3.6.2	5100173	4866958	50.8	43x	5135	97	22	6	4866958	213229	4	99.6	0.11
3.7.3	5286643	3492024	50.8	58x	5402	100	22	7	*Incomplete*	199250	4	99.29	1.91
3.7.11	4986415	4657310	50.8	46x	5046	93	22	7	4657310	214646	5	97.74	0.11
3.7.16	5039937	4807305	50.8	54x	5067	98	22	7	4807305	214536	5	99.6	0.11
3.7.25	4867813	1876384	51	36x	4896	90	22	16	*Incomplete*	-	8	98.05	3.5
3.7.28	5138354	4806102	50.7	61x	5177	98	22	6	4806102	214666	4	99.6	0.58
3.7.29.1.2	5156461	4824935	50.8	86x	5182	96	22	9	4824935	214579	7	99.6	0.11
3.7.30.1	5122274	4884953	50.8	42x	5168	100	22	6	4884953	216152	4	99.6	0.11
3.7.30.4	5005244	4852398	51	46x	5047	91	22	9	4852398	-	8	99.29	0.11
3.7.30.4.1	5260132	4933886	50.7	58x	5322	91	22	8	4933886	214700	6	99.6	0.11

In addition to the chromosome and pINV, each genome also harboured additional smaller plasmids (between 1-8 per genome, ranging from 1,459 to 108,503 bp). Plasmid typing using the MobTyper function from MobSUITE revealed a diverse plasmid repertoire within the isolate collection (presented in Table 4.2). Four types of Col plasmid (colicinogenic plasmids which encode the genes required for colicin production (Calcuttawala et al., 2017)) were identified across the collection: MOB_F/ColE1-like, Col (BS512), Col156 and Col (MG828) and together, they represented the second most frequently detected group of plasmids (after pINV). Between two and four Col plasmids were identified in every lineage 3 genome, and the combination of Col plasmids carried did not seem to correlate with clade or subclade. Except for lineage 2.8 (which harboured a MOB_F/ColE1-like and a Col(BS512) plasmid), Col plasmids were otherwise restricted to lineage 3 isolates.

Genomes were also screened for the presence of colicin genes independently of plasmid typing and interestingly, colicins were found to be encoded on plasmids that were not identified as Col plasmids by replicon type (Table 4.3), suggesting that plasmid typing alone is not predictive of actual colicin carriage in this case. From this analysis, genes encoding for a diverse range of colicin proteins were identified (with Colicins E1 and E3 most frequently detected) and colicins were found in almost all genomes, with no clear lineage-associated trend. A previous study highlighted an association between colicin carriage and the epidemiological success of some lineage 3 subclades (De Silva et al., 2023), however, this study did not include representatives of lineages 1 or 2. These results could suggest that colicins are important for driving sub-clade epidemiological success, but there are likely additional factors driving differences in epidemiological success between lineages 1, 2 and 3. All other plasmids had similar sequences to those previously reported in other Enterobacteriaceae, and no clear lineage-dependent trends were observed in the carriage of other plasmids, which were distributed sporadically throughout the collection.

Table 4.2. Summary of plasmid content in *S. sonnei* assemblies presented here. Plasmids were replicon and MOB typed using the MobTyper function in MobSUITE. The top line of each row indicates plasmid size, and the bottom row indicates the accession of the nearest Mash neighbour. For MOB_P and MOB_Q, no replicon types were detected, and for 'No-MOB' no replicon or MOB-types were identified. Plasmids of the completed genome of lineage 2.8 are included here for comparative purposes. * Denotes incomplete assemblies. The first column (IncFIA/IncFIC) represents pINV.

Genotype	IncFIA/IncFIC	MOB_F/Col-E1-like	Col(BS512)	Col156/MOB_Q	Col(MG828)	IncI2/MO BP	IncFIA,IncFII/ MOB_F,MO BP	IncK2/Z /MO BP	IncI1/B/O /MO BP	IncFIB	IncX1/MO BP	IncI-gamma/K1 /MO BP	MO BP	MOB_Q	No-MOB
Undetected (1)	blaTEM-1	-	-	-	-	aph 43917 bp ***LT985310***	-	-	102906 bp ***CP041565***	-	-	-	6888 bp ***CP014198***	11156 bp ***CP023649***	-
1.5	238313 bp ***CP000039***	-	-	-	-	-	125562 bp ***CP001065***	-	-	-	-	-	-	7084 bp ***CP023649***	-
2.1	216858 bp ***HE616529***	-	-	-	-	-	-	-	-	-	-	-	-	4074 bp ***CP018208***	3257 bp ***CP019023***
2.3	220157 bp ***HE616529***	-	-	-	-	-	-	-	-	-	-	-	5153 bp ***HE616530***	-	-
2.8	215774 bp ***HE616529***	5153 bp ***CP019897***	2089 bp ***HE616531***	-	-	-	-	-	-	-	-	-	-	-	8953 bp ***HE616532***
2.12.4	223662 bp ***CP023646***	-	-	-	-	-	77123 bp ***AP014877***	-	-	-	-	-	5153 bp ***HE616530***	-	-
3.4.1	215149 bp ***CP023646***	-	2088 bp ***CP033398***	5114 bp ***CP019693***	-	-	-	-	-	-	-	-	2717 bp ***CP038000***	6750 bp ***DQ916145***	2699 bp ***CP039609***
3.6.1	214316 bp ***CP023646***	2690 bp ***CP038000***	-	6015 bp ***KP970685*** 5114 bp ***CP019693***	1549 bp ***CP003038***	-	-	-	-	-	-	-	7939 bp ***KU932034***	4076 bp ***CP011140***	8401 bp ***CP034068***
3.6.1.1.1	212976 bp ***CP023646***	-	2089 bp ***CP030115***	-	-	-	-	-	-	-	-	-	2690 bp ***CP038000***	4878 bp ***KP970685*** 4269 bp ***CP019693***	8401 bp ***CP034068*** 3787 bp ***CP019138*** 2651 bp ***KX618698***
3.6.2	213229 bp ***CP023646***	2690 bp ***CP038000***	2089 bp ***CP030115***	5114 bp ***CP019693***	-	-	-	-	-	-	-	-	-	-	10093 bp ***CP034068***
3.7.3*	199250 bp ***CP023646***	2690 bp ***CP038000***	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.11	214646 bp ***CP016533***	2690 bp ***CP038000***	2101 bp ***CP023648***	5114 bp ***CP019693***	-	-	-	-	-	-	-	97999 bp ***CP016533***	6555 bp ***NC_008488***	-	-
3.7.16	214536 bp ***CP023646***	2690 bp ***CP038000***	2089 bp ***CP023648***	5114 bp ***CP019693***	1459 bp ***CP023264***	-	-	-	-	-	-	-	-	6744 bp ***KY348421***	-
3.7.25*	-	-	2101 bp ***CP023648***	5114 bp ***CP019693***	-	-	-	-	-	-	-	-	-	-	-
3.7.28	214666 bp ***CP023646***	5153 bp ***CP019897*** 2690 bp ***CP019897***	-	-	-	-	-	-	-	108503 bp ***CP034066***	-	-	1240 bp ***CP018984***	-	-
3.7.29.1.2	214579 bp ***CP023646***	2690 bp ***CP023653*** -	2101 bp ***CP023648***	-	1549 bp ***CP003038***	-	-	96229 bp ***CP026855***	-	-	-	-	-	6774 bp ***NC_022585***	4869 bp ***CP030008*** 2735 bp ***NC_011405***
3.7.30.1	216152 bp ***CP023646***	5153 bp ***CP023650*** 2690 bp ***CP038000***	-	-	-	-	-	-	-	-	-	-	-	-	8212 bp ***CP025235***
3.7.30.4	-	2687 bp ***CP038000***	2101 bp ***CP023648***	5166 bp ***CP034960***	-	-	-	-	82012 bp ***CP023387***	-	-	-	-	4082 bp ***KT693144*** 4072 bp ***CP012629***	7400 bp ***CP019281***
3.7.30.4.1	214700 bp ***CP023646***	7717 bp ***CP023650***	2101 bp ***CP023648***	5114 bp ***CP019693***	-	-	-	57356 bp ***CP018984***	-	-	35184 bp ***CP002111***	-	-	4074 bp ***LT985255***	

Table 4.3. Summary of plasmid encoded colicins. Colicins were identified using ABRicate and the full colicin database created previously (De Silva et al., 2023). This table directly corresponds to table 4.2. If annotated, genes are indicated on the top row, with the protein product in brackets. * Denotes incomplete assemblies. The first column (IncFIA/IncFIC) represents pINV.

Genotype	IncFIA/ IncFIC	MOB_F/Col-E1-like	Col(BS512)	Col156/ MOB_Q	Col(MG828)	IncI2/ MOB_P	IncFIA,IncFII/ MOB_F,MOB_P	IncK2/Z /MOB_P	IncI1/B/O /MOB_P	IncFIB	IncX1/ MOB_P	IncI-gamma/K1 / MOB_P	MOB_P	MOB_Q	No-MOB
Undetected (1)	-	-	-	-	-	- -	-	-	-	-	-	-	*cea* (Colicin-E1)	*ceaC* (Colicin-E3)	-
1.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2.1	-	-	-	-	-	-	-	-	-	-	-	-	-	*ceaC* (Colicin-E3)	-
2.3	-	-	-	-	-	-	-	-	-	-	-	-	-	*ceaC* (Colicin-E3)	-
2.8	-	*cea* (Colicin-E1)	-	-	-	-	-	-	-	-	-	-	-	-	-
2.12.4	-	-	-	-	-	-		-	-	-	-	-	-	*ceaC* (Colicin-E3)	-
3.4.1	-	-	-	-	-	-	-	-	-	-	-	-	-	*col* (Colicin-E9)	-
3.6.1	-	-	-	(Colicin-E3) -	-	-	-	-	-	-	-	-	-	-	-
3.6.1.1.1	-	-	-	-	-	-	-	-	-	-	-	-	-	(Colicin-E3) -	- -
3.6.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.3*	-	(Colicin-E1)	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.11	-	-	-	-	-	-	-	-	-	-	-	*cia* (Colicin Ia)	-	-	-
3.7.16	-	-	-	-	-	-	-	-	-	-	-	-	-	*col* (Colicin-E9)	-
3.7.25*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.28	-	*cea* (Colicin-E1) -	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.29.1.2	-	- -	-	-	-	-	-	-	-	-	-	-	-	(Colicin-E5)	- -
3.7.30.1	-	*cea* (Colicin-E1) -	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.30.4	-	-	-	-	-	-	-	-	-	-	-	-	-	- -	-
3.7.30.4.1	-	(Colicin 10)	-	-	-	-	-	-	-	-	-	-	-	-	-

Screening for AMR determinants confirmed that the selected isolates were representative of the AMR determinants that had previously been associated with the genotypes they were selected to represent (Table 4.4). Isolates belonging to clade 3.6.1 were both found to carry the *gyrA*_S83L mutation associated with this sub-clade (Chung The and Baker, 2018); the isolate representing subclade 3.6.1.1.1 also carried the additional QRDR mutations *gyrA*_D87G and *parC*_S80I, which are associated with the high-level fluoroquinolone resistance previously reported in subclade 3.6.1.1 (Chung The et al., 2019, Hawkey et al., 2021). All clade 3.6 isolates carried Tn7 harbouring *dfrA1* and *sat2* in the integron cassette (conferring resistance to trimethoprim and streptothricin), which was inserted between an IS4 family transposase and *glmS*. Similarly, all clade 3.7 isolates carried a distinct Tn7 variant harbouring *dfrA1, sat2* and *aadA1* (conferring additional aminoglycoside resistance) and the insertion occurred at the same chromosomal locus. Additionally, all isolates belonging to clade 3.6 were found to carry the small spA plasmid, encoding *tetA*, *aph(6)-Id, aph(3")-Ib* and *sul2* (Table 4.5), as is commonplace within this clade (Chung The et al., 2016). Genotypes 2.1 and 3.4.1 were both found to harbour the chromosomally encoded SRL which encodes for *aadA1*, *tetB*, *catA1* and *bla*OXA-1, typical of Latin American associated *S. sonnei* (Baker et al., 2018a).

Until now, the location of some resistance determinants as either chromosomally encoded (and so stably fixed) or plasmid encoded (which may be lost), was not entirely clear due to a lack of completed genomes. These results therefore confirm the previous conclusion that the acquisition of the Tn7 / Int2 pair is stably fixed in the chromosomes of clades 3.6 and 3.7 and highlight the insertion site to be the same despite the carriage of distinct variants. Moreover, resolving the location of resistance determinants may be beneficial for future use in experimental work, where stable chromosomally encoded genes may act as selective markers.

Table 4.4. Summary of antimicrobial resistance determinants identified by AMRFinder. † Denotes incomplete genome assemblies, * denotes genomes which are missing the virulence plasmid (pINV). Determinants in black were found to be chromosomally encoded, whilst those in red were found to be plasmid encoded.

	Antibiotic class							
Genotype	**BETA-LACTAM**	**AMINOGLYCOSIDE**	**TRIMETHOPRIM**	**SULFONAMIDE**	**PHENICOL**	**TETRACYCLINE**	**STREPTOTHRICIN**	**QUINOLONE**
1	*blaTEM-1*	*aph(6)-Id, aph(3")-Ib*	*dfrA14*	*sul2*				
1.5								
2.1	*blaOXA-1*	*aadA1*			*catA1*	*tet(B)*		
2.3								
2.8								
2.12.4	*blaTEM-1*	*aadA1, aph(6)-Id, aph(3")-Ib*	*dfrA1*	*sul2*		*tet(B)*	*sat2*	
3.4.1	*blaOXA-1*	*aadA1*			*catA1*	*tet(B)*		*qnrB19*
3.6.1		*aph(6)-Id, aph(3")-Ib*	*dfrA1*	*sul2*		*tet(A)*	*sat2*	*gyrA_S83L*
3.6.1.1.1		*aph(6)-Id, aph(3")-Ib*	*dfrA1*	*sul2*		*tet(A)*	*sat2*	*gyrA_S83L, gyrA_D87G, parC_S80I*
3.6.2		*aph(6)-Id, aph(3")-Ib*	*dfrA1*	*sul2*		*tet(A)*	*sat2*	
3.7.3 †		*aadA1*	*dfrA1*				*sat2*	
3.7.11		*aadA1*	*dfrA1*				*sat2*	
3.7.16		*aadA1*	*dfrA1*				*sat2*	
3.7.25 †*		*aadA1*	*dfrA1*			*tet(B)*	*sat2*	
3.7.28		*aadA1*	*dfrA1, dfrA51*				*sat2*	
3.7.29.1.2		*aadA1, aph(6)-Id, aph(3")-Ib*	*dfrA1*	*sul2*			*sat2*	*gyrA_S83L*
3.7.30.1		*aadA1*	*dfrA1*				*sat2*	
3.7.30.4 *		*aadA1*	*dfrA1*				*sat2*	
3.7.30.4.1		*aadA1*	*dfrA1*				*sat2*	

Table 4.5. Summary of plasmid encoded antimicrobial resistance determinants. This table corresponds directly to Table 4.2. † Denotes incomplete genome assemblies, * denotes genomes which are missing the virulence plasmid (pINV).

Genotype	IncFIA/IncFIC	MOB_F/Col-E1-like	Col(BS512)	Col156/MOB_Q	Col(MG828)	IncI2/MOBP	IncFIA,IncFII/MOB_F,MOBP	IncK2/Z/MOBP	IncI1/B/O/MOBP	IncFIB	IncX1/MOBP	IncI-gamma/K1/MOBP	MOBP	MOB_Q	No-MOB
Undetected (1)	*blaTEM-1*	-	-	-	-	*blaTEM-1* *blaTEM-1*	-	-	*aph(6)-Id* *blaTEM-1* *dfrA14* *aph(3'')-Ib* *sul2*	-	-	-	-	-	-
1.5	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2.3	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2.8	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
2.12.4	-	-	-	-	-	-	*aph(6)-Id* *blaTEM-1* *aph(3'')-Ib* *sul2* *tet(B)*	-	-	-	-	-	-	-	-
3.4.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*qnrB19*
3.6.1	-	-	-	- -	-	-	-	-	-	-	-	-	-	-	*aph(6)-Id* *aph(3'')-Ib* *sul2* *tet(A)*
3.6.1.1.1	-	-	-	-	-	-	-	-	-	-	-	-	-	- -	*aph(6)-Id* *aph(3'')-Ib* *sul2* *tet(A)* - -
3.6.2	-	-	-	-	-	-	-	-	-	-	-	-	-	-	*aph(6)-Id* *aph(3'')-Ib* *sul2* *tet(A)*
3.7.3*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.11	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.16	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.25*	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.28	-	- -	-	-	-	-	-	-	-	*dfrA51*	-	-	-	-	-
3.7.29.1.2	-	- -	-	-	-	-	-	*aph(6)-Id* *aph(3'')-Ib* *sul2*	-	-	-	-	-	-	- -
3.7.30.1	-	- -	-	-	-	-	-	-	-	-	-	-	-	-	-
3.7.30.4	-	-	-	-	-	-	-	-	-	-	-	-	-	- -	-
3.7.30.4.1	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-

Genomes were next screened for the presence of virulence associated genes (VAGs) using VFDB to identify any lineage dependent trends in virulence. The greatest number of VAGs were identified in lineages 1 and 2 (89-91), whilst complete lineage 3 genomes harboured the fewest (78-80); two lineage 3 isolates harboured only 27 VAGs, but these isolates were both missing pINV (Table 4.6). The most frequently detected VAGs were T3SS effectors or T3SS associated genes, with 54-56 genes present in every pINV-positive genome. Iron acquisition genes were conserved throughout the collection, with a total of 18 detected in every genome; the same was true for the 'other' category (which included *ompA,* an outer membrane protein and *senB* and *sigA,* which are both enterotoxins*),* where all three genes were present in every genome analysed. The most variation was observed in the detection of adhesins and fimbriae, with 12-14 detected in each lineage 1 and 2 genome, but only three present in every lineage 3 genome: *csgB, csgF* and *csgG*. These genes are all involved in the biogenesis of curli; thin, aggregative fimbriae implicated in adhesion and biofilm formation and are widely conserved in *E. coli* and *Salmonella* (Barnhart and Chapman, 2006, Zakikhany et al., 2010). *csgD,* which encodes for the master regulator of curli biogenesis (Ogasawara et al., 2011, Zakikhany et al., 2010), was restricted to only lineage 1 genomes, suggesting that this operon is undergoing degradation in lineages 2 and 3. *fimBHGF*, which encodes for the type 1 fimbriae (Klemm, 1986) and *faeCDEFHI*, which encodes for the ETEC-associated K88 fimbriae (Mol et al., 2001), were both present in most lineage 1 and 2 genomes (the *fae gene* cluster was also absent from lineage 1.5) but absent from all lineage 3 genomes, again suggestive of adaptive loss. Overall, the data presented here are suggestive of pathovariant formation within lineage 3, characterised by a reduction in both chromosome and pINV sizes and gives an insight into which processes (such as the loss of immunogenic features) may be under selection in the evolution of *S. sonnei*.

Table 4.6. Virulence associated genes (VAGs) detected by VFDB. The 'other' category represents outer membrane protein *ompA* and enterotoxins *sigA and senB*. † Denotes incomplete genome assemblies, * denotes genomes missing the virulence plasmid. (pINV).

Genotype	Adhesins and fimbriae	Iron acquisition	T3SS associated	Other	Total
Undetected (1)	14	18	54	3	90
1.5	12	18	55	3	89
2.1	13	18	56	3	91
2.3	13	18	56	3	91
2.8	13	18	56	3	91
2.12.4	13	18	55	3	90
3.4.1	3	18	55	3	80
3.6.1	3	18	55	3	80
3.6.1.1.1	3	18	54	3	80
3.6.2	3	18	55	3	80
3.7.3 †	3	18	53	3	78
3.7.11	3	18	55	3	80
3.7.16	3	18	55	3	80
3.7.25 †*	3	18	3	3	27
3.7.28	3	18	55	3	80
3.7.29.1.2	3	18	55	3	80
3.7.30.1	3	18	55	3	80
3.7.30.4 *	3	18	3	3	27
3.7.30.4.1	3	18	55	3	80

4.2.3. Pangenome analysis reveals limited gene content variation in *S. sonnei*

To explore overall gene content variation within *S. sonnei* lineages, a pangenome (including only complete, pINV-positive genomes) was next built using Panaroo. The pangenome encompasses both the core genome, which consists of genes shared across all isolates, and the accessory genome, genes that are found in some, but not all genomes. A total of 6317 homologous gene clusters (HGCs) were identified from the 16 complete and pINV-positive *S. sonnei* genomes analysed: of these, 4282 genes (67.78%) were present in > 99% of strains, representing the core genome; 1268 genes (19.83%) were present in ≥ 15% of strains, representing 'shell' genes and 767 genes (12.12%) were present in < 15%, representing the 'cloud' genes (Fig 4.5).

The number of unique HGCs per isolate varied from 1-104 (Table 4.7). Lineage 1 had the most lineage-specific HGCs (HGCs present in every isolate of that lineage, and absent in the other lineages) at 63. Lineage 2 had only 6 lineage-specific HGCs and lineage 3 had 27 lineage-specific HGCs. Of the lineage specific HGCs in lineages 2 and 3, many were annotated as ISs or IS accessory genes. Since there were multiple isolates from clades 3.6 and 3.7, clade-specific genes were also identified for these clades. For clade 3.6, there were 10 specific HGCs and for clade 3.7 there were no clade-specific HGCs identified in the genomes analysed. Since both clade 3.6 and clade 3.7 are together responsible for the greatest burden of contemporary infections, I also looked for HGCs that occurred only within these clades but only one HGC appeared in every clade 3.6 and 3.7 isolate, which was annotated as an IS4 family transposase.

To determine if lineage specific HGCs were associated with a particular biological process, genes were classified using the PANTHER GO-slim tool. For lineage 1, most HGCs could not be classified into a biological processes category, but of the 20 that were, the most common categories were cellular processes and metabolic processes (Table 4.8), concomitant with previous findings suggesting a reduced metabolic capacity in lineage 3 (Holt et al., 2012). For lineage 2 and 3, most lineage specific HGCs were annotated as ISs or IS associated genes, meaning they were not assigned into a PANTHER GO-biological process category. Overall, these results highlight minor differences in unique gene content in *S. sonnei* lineages, with the majority of genes fitting into the core gene category. Where there were differences in gene content, these seemed to be sporadically distributed throughout lineages (since there were very few lineage specific HGCs) and on most occasions, unique HGCs were either not annotated as known genes or were annotated as ISs, suggesting that variation in *S. sonnei* is not largely driven by the binary presence or absence of genes.

Figure 4.5. Linear visualisation of the *S. sonnei* pangenome plotted alongside a maximum-likelihood phylogenetic tree. Panaroo was used to build a pangenome and a core genome alignment. FastTree was used to build a maximum likelihood tree, which is rooted at the midpoint. The resulting outputs were then visualised using Phandango. Blue stripes indicate the presence of a gene, and white stripes indicate the absence of a gene. L = lineage. Core genes are those present in > 95% of genomes, shell genes are defined by presence in ≥ 15% of genomes and cloud genes are those present in < 15% of genomes.

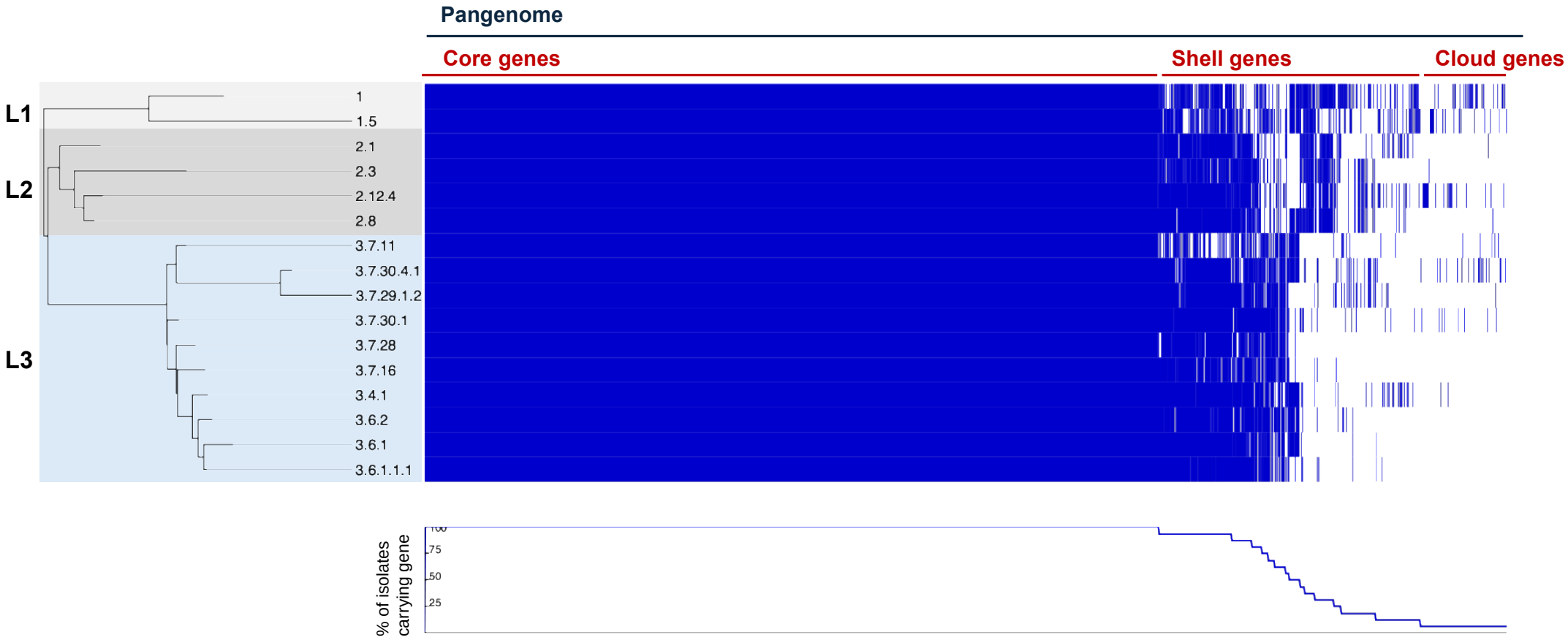


Table 4.7. The number of unique homologous gene clusters (HGCs) to each isolate and to each lineage. Genes were annotated using Bakta and HGCs were identified using Panaroo.

Lineage	No. genes unique to isolate	No. genes unique to lineage
1	95	63
1.5	104	
2.1	9	6
2.3	18	
2.8	6	
2.12.4	78	
3.4.1	24	27
3.6.1	5	
3.6.1.1.1	3	
3.6.2	3	
3.7.11	3	
3.7.16	3	
3.7.28	1	
3.7.29.1.2	11	
3.7.30.1	31	
3.7.30.4.1	90	

Table 4.8. Classification of lineage specific HGCs by Biological Processes gene ontology (GO) categories. Genes were classified using the PANTHER GO-slim tool.

PANTHER GO-Biological Processes Category	Lineage 1 HGCs	Lineage 2 HGCs	Lineage 3 HGCs
Unclassified	43	6	27
Metabolic processes (GO:0008152)	8	-	-
Cellular processes (GO:0009987)	8	-	-
Localisation (GO:0051179)	1	-	-
Biological regulation (GO:0065007)	3	-	-

4.2.4. Variations in the burden of insertion sequences (ISs) in *S. sonnei* lineages

Previous studies have produced evidence for ongoing IS activity within *S. sonnei*, with data from short-read sequences highlighting the proliferation of ISs specifically within lineages 2 and 3 (Hawkey et al., 2020), but this has not been examined in completed chromosomal sequences, and never in pINV due to a lack of genomes containing pINV. The abundance and composition of ISs across *S. sonnei* lineages with complete genomes were therefore analysed next. A total of 16 distinct IS families were identified by ISEscanner across all genomes (Fig 4.6 A). There was some variation in the total IS burden by lineage, with lineage 1 genomes harbouring the fewest, with an average of 480, lineage 2 an average of 490 whilst lineage 3 genomes had an average of 501. Most ISs were identified across all lineages, except for IS*30* which was only detected in lineage 1.5 as a single copy encoded on neither the chromosome nor pINV, but another smaller plasmid.

The detected ISs were split based on their presence in the chromosome, pINV or on other plasmids to account for their different evolutionary histories and different evolutionary dynamics. Considering the abundance of ISs in the chromosome alone revealed a clearer trend, with lineage 1 harbouring the fewest chromosomal ISs (an average of 385) and lineage 3 harbouring the most (an average of 426.3) (Fig 4.6 B). There was also some variation in the contribution of each IS family to the total burden of ISs between lineages (Table 4.9). Five families (IS1, IS3, IS21, IS4 and IS110) accounted for > 94% of the total proportion of IS in all three lineages, whilst the other nine IS families were detected at much lower frequencies. IS3, IS4, ISAS1, IS630 and IS256 all contributed to a greater IS proportion in lineage 3 as compared to lineage 1, whilst other IS families declined in frequency or did not change.

In pINV, a total of 12 different IS families were identified and two of these were unique to pINV: IS5 and ISL3, neither of which have been reported in *S. sonnei* before, likely owing to a lack of pINV sequences. Interestingly, a contradictory trend in IS abundance in pINV (compared to chromosomal IS abundance) was observed (Fig 4.6 C). Although the differences were smaller; an average of 80 ISs were detected in lineage 1; an average of 71.5 was detected in lineage

2 and an average of 73.2 were detected in lineage 3 genomes. IS630 and IS91 both contributed to a greater proportion of IS burden in lineage 3 when compared to lineage 1, whilst all other families decreased or remained the same (Table 4.10). For ISs carried on other plasmids, lineage 1.5 harboured the most (25), which is consistent with its carriage of the largest plasmid (Table 4.2). ISs carried on other plasmids were sporadically detected in single strains, but there was no clear lineage-dependent pattern (Fig 4.6 D), in line with there being limited lineage-dependent trends in plasmid carriage.

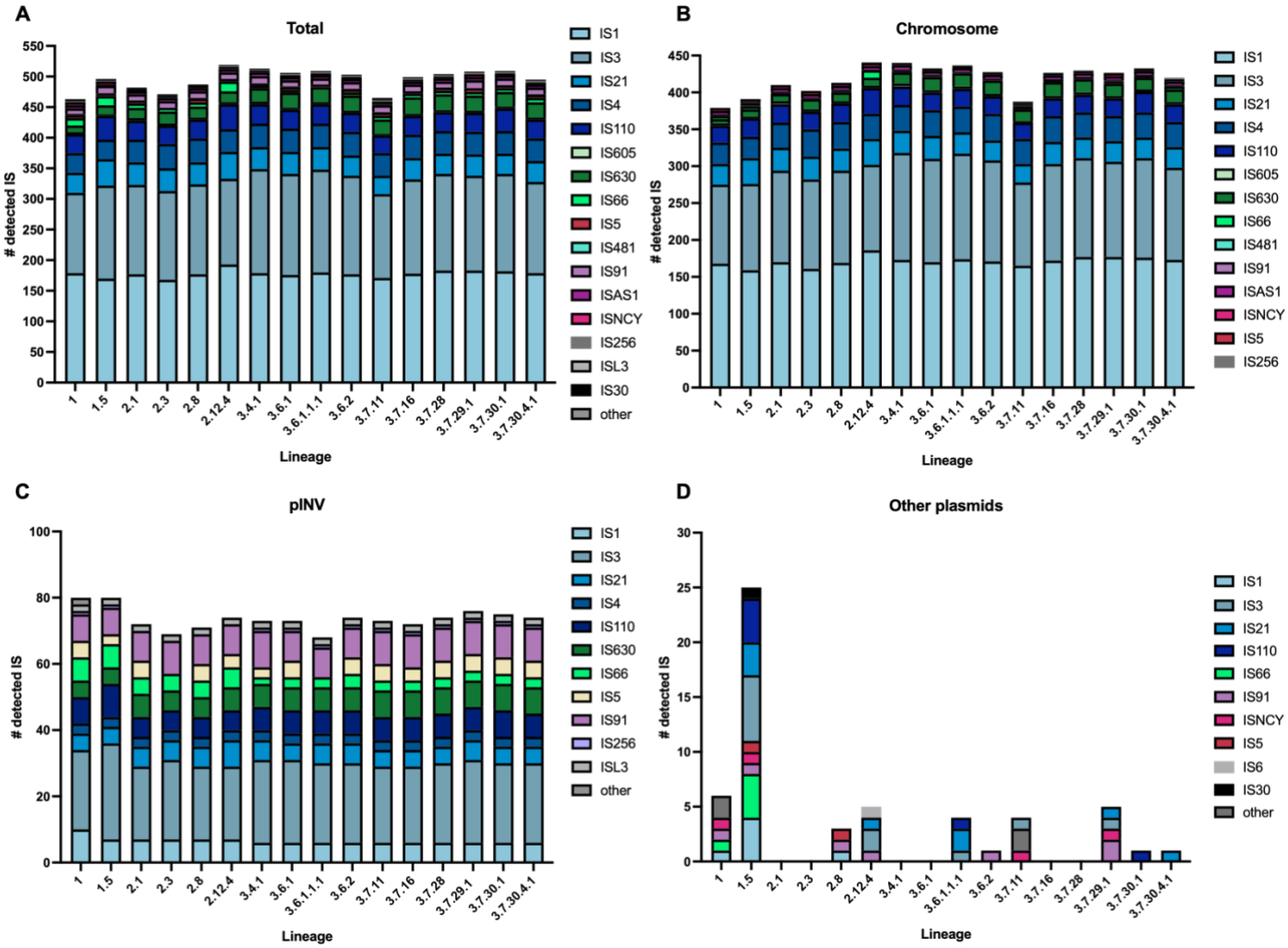


Figure 4.6. Insertion sequences (ISs) detected in complete *S. sonnei* genome sequences by ISEscanner. A) Total number of ISs identified across all contigs of completed genomes. **B)** ISs identified in completed chromosomal sequences. **C)** ISs identified in pINV sequences **D)** ISs identified in other plasmids present in *S. sonnei* genomes. ISs were detected using ISEscanner, 'other' represents ISs that could not be matched to a known IS family within the database used.

Table 4.9. Average proportion of each insertion sequence (IS) family in the chromosomes of *S. sonnei* genomes. IS were identified using ISEscanner. Lineage 1 (n = 2), Lineage 2 (n = 4), lineage 3 (n = 10).

	Proportion of chromosomal IS (%)		
	Lineage 1	Lineage 2	Lineage 3
IS1	42.50	41.15	40.56
IS3	29.08	29.23	31.21
IS21	8.17	7.62	6.66
IS4	7.53	8.48	8.12
IS110	6.23	6.51	5.70
IS605	0.65	0.78	0.70
IS630	2.07	2.83	4.04
IS66	1.04	0.87	0.57
IS5	0.00	0.00	0.02
IS481	0.39	0.06	0.12
IS91	0.38	0.24	0.21
ISAS1	0.92	1.20	1.12
ISNCY	1.04	0.96	0.73
IS256	0.00	0.06	0.24

Table 4.10. Average proportion of each insertion sequence (IS) family in the virulence plasmid (pINV) of *S. sonnei* genomes. ISs were identified using ISEscanner. Other represents ISs that were not assigned a known family. Lineage 1 (n = 2), Lineage 2 (n = 4), lineage 3 (n = 10).

	Proportion of pINV IS (%)		
	Lineage 1	Lineage 2	Lineage 3
IS1	10.63	9.80	8.20
IS3	33.13	31.51	32.94
IS21	6.25	9.07	7.39
IS4	3.75	4.20	4.10
IS110	11.25	8.40	9.70
IS630	6.25	9.08	10.38
IS66	8.75	7.34	4.10
IS5	5.00	4.85	5.69
IS91	10.00	12.96	13.39
IS256	1.25	0.00	1.37
ISL3	2.50	2.80	2.73
other	1.25	0.00	0.00

The accumulation and proliferation of ISs have been linked to an increase in gene pseudogenisation (Siguier et al., 2014), so to investigate any changes in gene functionality, genomes were next screened for the presence of pseudogenes using annotations performed by PGAP. The number of pseudogenes predicted using this method is higher than previous predictions (for lineage 2.8 at least), but this is likely due to differences in the annotation method, whereby PGAP may be recording 2 separate pseudogenes in situations where a single gene is interrupted by an IS. In agreement with the increased abundance of ISs in lineage 3 isolates, this analysis also revealed a similar increase in the total number of pseudogenes present in lineage 3 genomes, with an average of 711 identified compared to 671 in lineage 2 and 663.5 in lineage 1 (Fig 4.7). Lineage 3 had a greater number of pseudogenes compared to both lineage 1 and lineage 2 and differences in pseudogene number were mostly concentrated in the chromosome. In pINV, the same pattern as IS abundance was observed in the number of pseudogenes, where lineage 1 genomes had the largest number of pseudogenes, likely due to the increased pINV size. Collectively, these results are indicative of ongoing IS activity in *S. sonnei*, with lineage dependent variation observed in both the abundance and the proportion of different IS families. The same trends were also reflected in the number of pseudogenes, which suggests that IS expansion is resulting in functional changes within *S. sonnei* genomes.

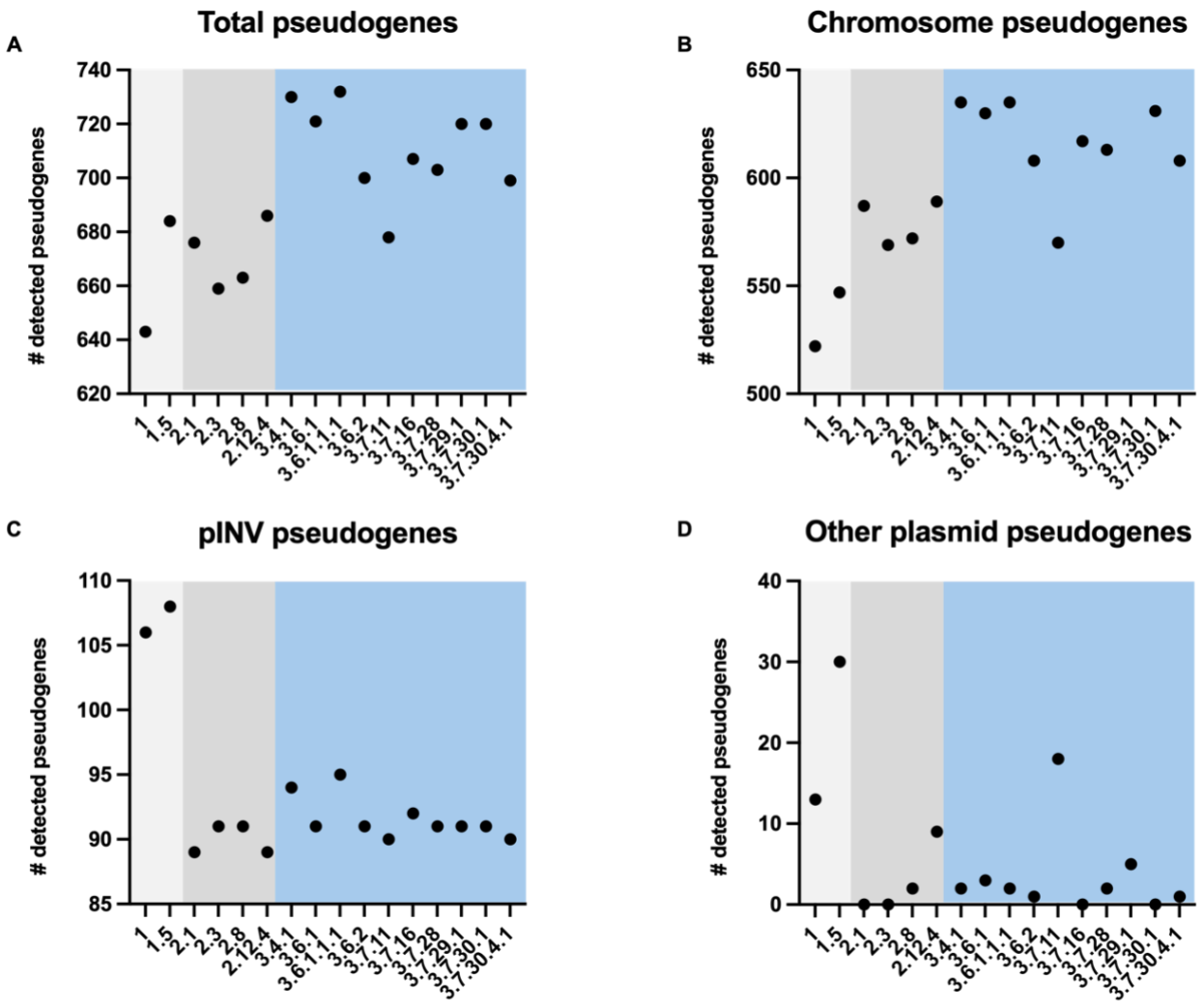


Figure 4.7. Abundance of pseudogenes predicted by the NCBI prokaryotic genome annotation pipeline (PGAP). A) Total pseudogenes identified across all contigs of completed genomes **B)** Pseudogenes identified in completed chromosomal sequences **C)** Pseudogenes identified in pINV sequences **D)** Pseudogenes identified in other plasmids present in *S. sonnei* genomes. Lineage 1 isolates are in light grey, lineage 2 in dark grey and lineage 3 in blue.

4.2.5. Variability in the structure of *S. sonnei* genomes

Genome rearrangements are key drivers of bacterial evolution, facilitating adaptation without directly modulating genome content, but instead modifying the structural arrangement of the genome, which can, in turn, alter gene expression patterns (Darling et al., 2008). *Shigella* genomes are documented to have an exceptionally high rate of rearrangements compared to other pathogenic (and non-pathogenic) *E. coli* (Seferbekova et al., 2021) likely mediated by the similar overrepresentation of ISs in their genomes (Hawkey et al., 2020). Owing to the variation in IS abundance and composition I identified within *S. sonnei* lineages, I next performed a whole genome alignment of chromosomal and pINV sequences to assess for any lineage-dependent structural variation.

Firstly, chromosomal sequences were aligned using progressiveMauve, which revealed the presence of several structural variations, some of which were conserved in a lineage-dependent manner (Fig 4.8).

Deletions

Four conserved large-scale deletion events (> 7 kbp) were identified in total, two of which were common to all lineage 2 and 3 genomes, and two of which were common to only lineage 3 genomes. Firstly, a region of ~22 kbp, situated between *xylA* and a hypothetical protein, was deleted in all lineage 2 and 3 genomes. The deletion included 21 annotated genes and one small RNA (Table 4.11). Many of the deleted genes had a role in catabolism, including *xylFGH* and *xylR*, which are essential for the utilisation of $_{D}$-xylose (Song and Park, 1997). The inability to ferment $_{D}$-xylose is a common feature of *Shigella* (Lan et al., 2004), mediated through the loss of *xylA* in *S. flexneri* and the loss of both *xylAB* and *xylFGH* in *S. dysenteriae* (Yang et al., 2005), suggesting that this deletion contributes to the formation of *S. sonnei* pathovariants.

Figure 4.8. Whole genome alignment of *S. sonnei* chromosomal sequences using progressiveMauve. Coloured blocks represent locally colinear blocks of sequence homology, with respect to the reference genome (lineage 1.5 in this case, indicated by (R)). Inverted regions are shown on the bottom strand. Position 1 corresponds to *fabB* in all genomes. Blank regions indicate regions with a reduced average sequence homology. Numbers at the top indicate chromosomal position.

Table 4.11. Conserved ~22 kbp deletion in lineage 2 and 3 genomes (with reference to lineage 1). Gene product and functions included as predicted by Bakta.

Gene name(s)	Gene product /function	Additional notes
xylFGH, xylR	$_{D}$-xylose ABC transporter and transcriptional regulator	Essential for the utilisation and transport of $_{D}$-xylose in *Shigella* (Song and Park, 1997)
bax, baxL	Protein bax	Putative glycoside hydrolase (Morita et al., 2022)
malS	Alpha-amylase	Encodes for a periplasmic alpha amylase, important in maltodextrin metabolism in *Shigella* (Schneider et al., 1992)
avtA	Valine-pyruvate transaminase	Encodes a alanine-valine transaminase, essential for alanine synthesis in *E. coli* K-12 (Wang et al., 1987)
ES036	Small RNA	-
ysaA	Putative electron transport protein	Encodes a putative ferredoxin-like protein, involved in electron transport and redox sensing in *E. coli* K-12 (Pinske et al,. 2018)
yiaJKLMNO	2,3-diketo-L-gulonate transporter	Important for $_{L}$-xyulose utilisation in *E. coli* K-12 (Ibañez et al., 2000)
lyxK	L-xylulokinase	Encodes an $_{L}$-xyulose kinase in *E. coli* K-12, involved in the positive regulation of $_{L}$-xyulose synthesis (Pomposiello et al., 2001)
ulaD	3-keto-L-gulonate-6-phosphate decarboxylase	Involved in the anaerobic utilisation of $_{L}$-ascorbate in *E. coli* K-12 (Zhang et al., 2003)

sgbU	L-ribulose-5-phosphate 3-epimerase	Required for the utilisation of $_L$-lyxose in *E. coli* K-12 (Ibañez et al., 2000)
araD	L-ribulose-5-phosphate 4-epimerase	Involved in the utilisation of $_L$-arabinose (Better, 1999)
yesN	Two component response regulator	Regulation of carbohydrate utilisation in *Lactobacillus* (Xu et al., 2015)
melB	Na+/melibiose symporter or related transporter	Encodes for a melibiose carrier in *E. coli* K-12 (Yazyu et al., 1984)

Another deletion of ~7 kbp was observed in all lineage 2 and 3 genomes. The deleted region was located in between two hypothetical proteins and consisted of ten coding sequences: four ISs (three IS3 family and IS630 family), three hypothetical proteins and three genes with known functions (*yehL*, *yehM* and *zorO*). The *yehL* and *yehM* genes are both uncharacterised (but may belong to the putative *yeh* fimbrial cluster (Ravan and Amandadi, 2015)) and the *zorO* gene encodes for the toxin component of the *zorO-orzO* toxin-antitoxin system (Wen et al., 2014).

Observed only in lineage 3 genomes was a conserved deletion of ~16 kbp between *yahE* and *ykgG.* The deleted region included eleven CDSs, most of which do not have a well-defined function in *Shigella* (Table 4.12). Of particular note was the loss of *ehaA*, an autotransporter adhesin with a role in biofilm formation in Shiga toxin producing *E. coli* (STEC) (Wells et al., 2008). Data presented in section 4.2.2 demonstrated the conserved loss of adhesins and fimbriae in lineage 3 genomes, and the loss of *ehaA* observed here likely represents another example of convergent loss.

Table 4.12. Conserved 16 kbp deletion in lineage 3 genomes (with reference to lineage 1). Gene product and functions included as predicted by Bakta, the locus tag represents that assigned by Bakta in the lineage 1 genome. STEC = Shiga toxin producing *E. coli*.

Gene name(s)	Gene product /function	Additional notes
yahBCD	Uncharacterised proteins	*yahB* is a putative transcription factor with a predicted role in metabolism (Duarte-Velázquez et al., 2022)
pdeL	Cyclic di-GMP phosphodiesterase	Predicted regulator of *fli* operon (Yilmaz et al., 2020)
ehaA	Autotransporter adhesin	Contributes to adhesion and biofilm formation in STEC (Wells et al., 2008)
Hypothetical protein	-	-
betT, betI, betAB	Glycine betaine synthesis	Confers osmotic tolerance in *E. coli* (Landfald and Strøm, 1986)
Unnamed (locus tag IAHEGK_19900)	Inner membrane protein	-

Finally, a deletion of ~7 kbp, which was inclusive of the type 1 fimbrial genes identified as absent in section 4.2.2, was documented in all lineage 3 genomes. The deleted region was between *gntP* and a hypothetical protein and consisted of 2 hypothetical proteins, a copy of IS*Ss06* (IS110 family) and the *fimHGFDB* genes. This finding is consistent with a previous study which reported the deletion of the whole *fim* operon in a lineage 3 isolate of *S. sonnei* (Bravo et al., 2015). As discussed in section 4.2.2, the independent loss of type 1 fimbriae has been reported in all *Shigella* and EIEC subgroup, and here evidence for ongoing degradation in *S. sonnei* lineages is provided, suggesting that its presence is detrimental to *Shigella*.

Insertions

Conserved, large-scale insertion events were uncommon in lineage 2 and 3 genomes. Insertions were mostly attributed to the acquisition of lineage-specific chromosomally encoded AMR determinants, and phage integrations, which occurred independently of lineage. In all clade 3.6 and 3.7 genomes, the Int2 bearing Tn7 cassette was identified as part of a ~13 kbp insertion. The insertion occurred at the same site in both clade 3.6 and 3.7 genomes (adjacent to *glmS*) but the contents of the insertion differed slightly, with *aadA1* present in only clade 3.7 genomes (Fig 4.9 A). The presence of distinct Tn7/Int2 variants suggests that this insertion occurred independently in clades 3.6 and 3.7 prior to international spread and subsequent clonal expansion, consistent with previous findings (Holt et al., 2012).

The insertion of the SRL PAI was also detected in lineage 2.1 and 3.4.1 genomes. The insertions occurred at different sites in the chromosome, but in both cases inserted into a copy of *trnS* (which encodes for $tRNA^{Ser}$). In lineage 2.1 the insertion was ~70 kbp, whilst for lineage 3.4.1 it was larger at ~80 kbp. As well as the AMR determinants encoded on the SRL (introduced in section 4.2.2), the insertion also included the *fec* gene cluster (which encodes for the ferric dicitrate transport system (Luck et al., 2001)), several hypothetical proteins and genes involved in DNA mobility. Insertions were similar for both genomes, but the lineage 3.4.1 insertion included more hypothetical proteins and genes implicated in DNA mobility compared to lineage 2.1 (Fig 4.9 B). Acquisition of the SRL is widespread in many *Shigella* subgroups and previous work has highlighted its frequent insertion into *Shigella* tRNA genes, suggestive of a preferred integration site (Williams, 2002, Luck et al., 2001). The independent insertions of the SRL PAI into two separate *S. sonnei* lineages further implies that it confers a fitness advantage which has largely assumed to be antibiotic resistance (Fullá et al., 2005). However, the presence of an iron transport system, and many other currently uncharacterised proteins could confer additional advantages which currently remain poorly understood and should therefore be explored in future work.

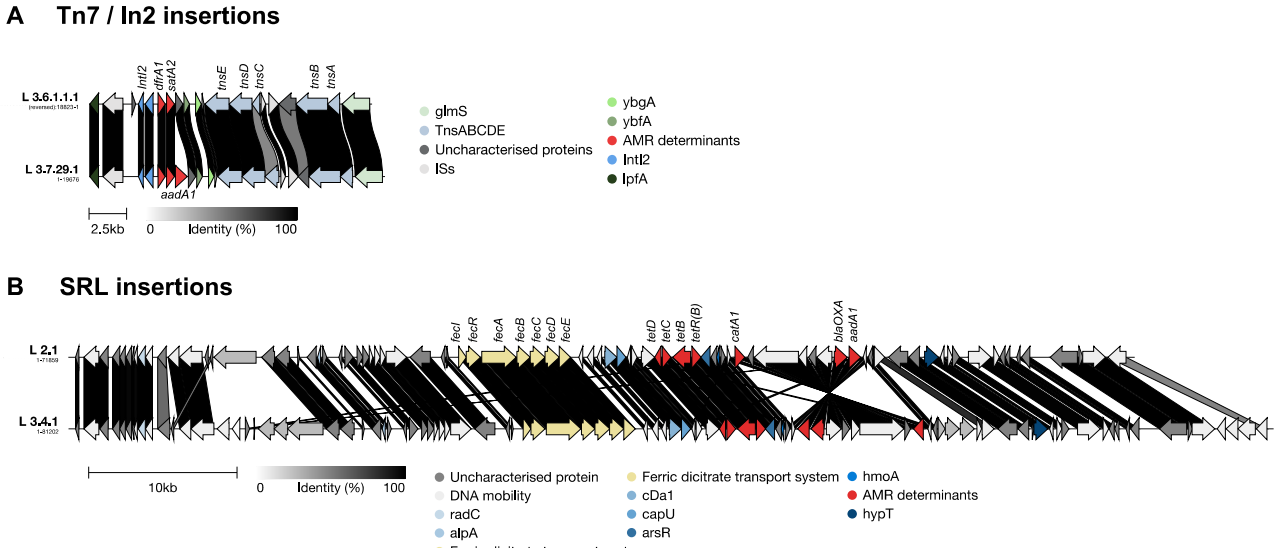


Figure 4.9. Clinker gene cluster comparison of Tn7 / Int2 and SRL insertions. A) Gene cluster comparison of the Tn7 / Int2 region in representative genomes of clades 3.6 and 3.7. **B)** Gene cluster comparison of the *Shigella* resistance loci (SRL) pathogenicity island in lineages 2.1 and 3.4.1. Homologous regions are linked through the black/grey bars which also indicate the percentage identity. AMR = antimicrobial resistance.

Inversions

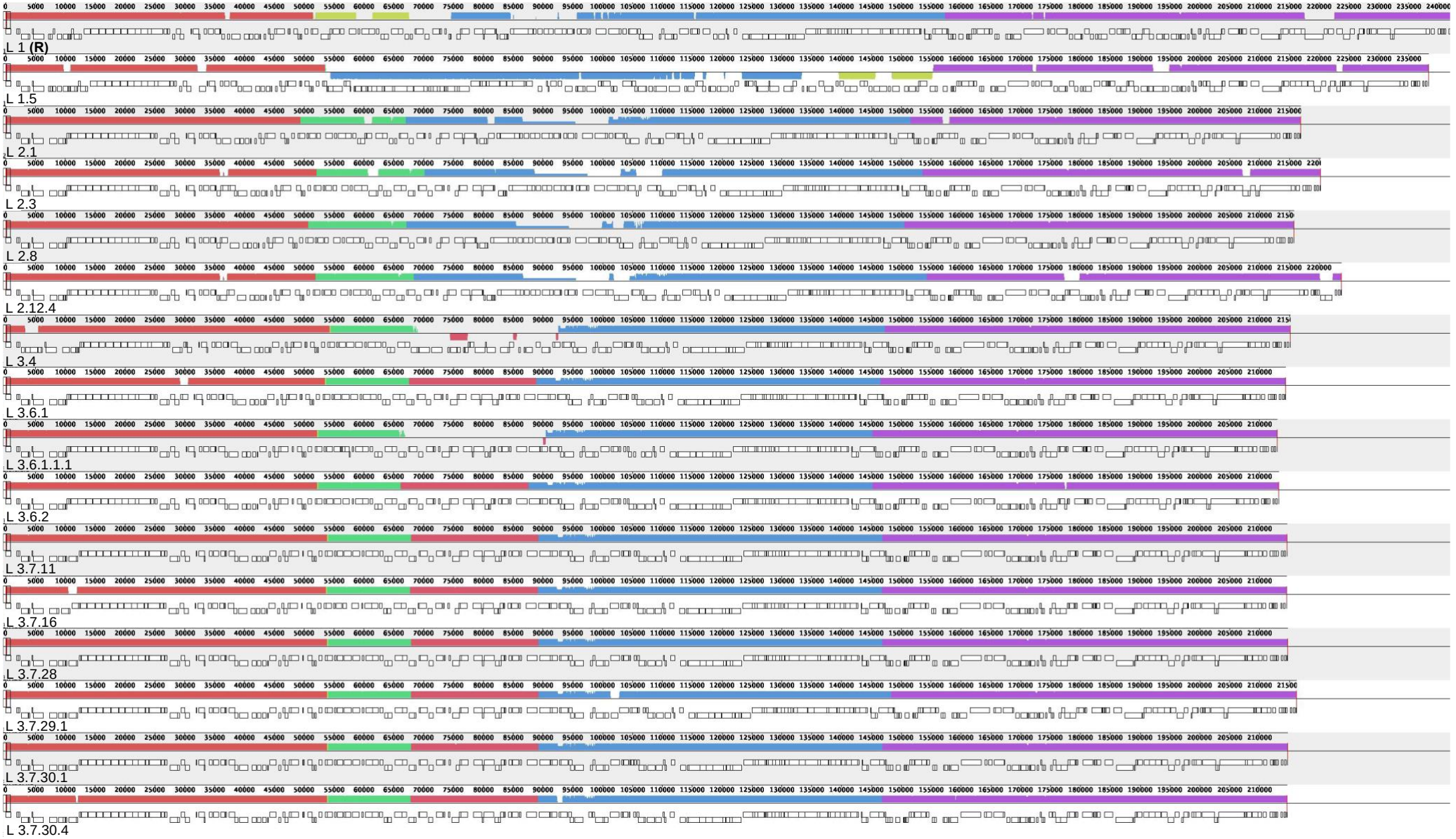
Two large-scale inversions were identified in all lineage 2 and 3 genomes: one inversion of ~20 kbp occurred adjacent to *idnO,* and a second inversion was identified adjacent to *bamD,* upstream of a 23s rRNA ribosomal subunit. In all lineage 2 genomes, the inverted region adjacent to *bamD* was ~45 kbp, whilst for all lineage 3 genomes, it presented as a translocation inversion of ~22 kbp. In both cases, the inversions adjacent to *bamD* were flanked by IS4 family transposases, suggestive of homologous recombination between ISs and highlighting evidence of ongoing IS activity.

I identified three large-scale inversions that were restricted to lineage 3 genomes. Firstly, an inversion of ~140 kbp was observed adjacent to *ssT*, but a region of ~93 kbp (flanked by IS*630* copies) appeared to have reversed back on to the leading strand. A second inversion adjacent to the *mhpC* gene was identified in lineage 3 genomes; in most cases, the inverted region was ~34 kbp and was flanked on both sides by IS3 family transposase but in lineage 3.6.2 genomes, a ~40 kbp Mu-like prophage had inserted in the middle of the inversion. In all cases, the inverted region contained *mhpR*, which acts as the transcriptional regulator for the *mhp* gene cluster (involved in the catabolism of small aromatic molecules (Xu and Zhou, 2020)). The *mhpR* gene was subsequently moved away from *mhpAB* which could have implications for gene expression. Interestingly, the *mhp* operon was previously highlighted as absent from *S. flexneri* 2457T (Wei et al., 2003)*,* implying that this metabolic pathway might be under convergent selection in other *Shigella* subgroups. Finally, rearrangements were observed adjacent to *rtcB* in all lineage 3 genomes, but the rearrangements varied and were not flanked by ISs. In clade 3.4 and 3.6 genomes, a ~130 kbp region was inverted, whilst it presented as a ~107 kbp translocation inversion in clade 3.7 genomes. The implications of these chromosomal rearrangements currently remain unclear, but potential impacts on gene expression, or inactivation, are likely to be important for conferring adaptive changes.

Having identified several chromosomal rearrangements, the alignment of pINV sequences was performed next, which revealed the presence of 6 LCBs, with limited large scale structural variation (Fig 4.10).

The variation observed previously in pINV size (shown in section 4.2.2) did not seem to be mediated by a single large deletion event, but instead by several smaller deletions. Such an example of a deletion event in pINV is an ~11 kbp region that was present in the lineage 1 genome and all lineage 2 genomes but absent in lineage 1.5 and all lineage 3 genomes. The deleted region included the *fae* gene cluster, which was previously identified as absent in section 4.2.2. The deleted region was flanked by a copy of an IS3 family transposase on one side and *araC* on the other side (Fig 4.11). Although this deletion could not be directly linked to a recombination event between 2 flanking homologous IS copies, the large abundance of IS in the genomic region is indicative of ongoing IS activity, which could also play a role in the regulation of T3SS related virulence, given its proximity to the T3SS master regulator *virF*.

Figure 4.10. Whole genome alignment of *S. sonnei* pINV sequences using progressiveMauve. Coloured blocks represent locally colinear blocks of sequence homology, with respect to the reference genome (lineage 1 in this case, indicated by (R)). Position 1 corresponds to *repA* in all genomes. Inverted regions are shown on the bottom strand. Blank regions indicate regions with an overall reduced average sequence homology.



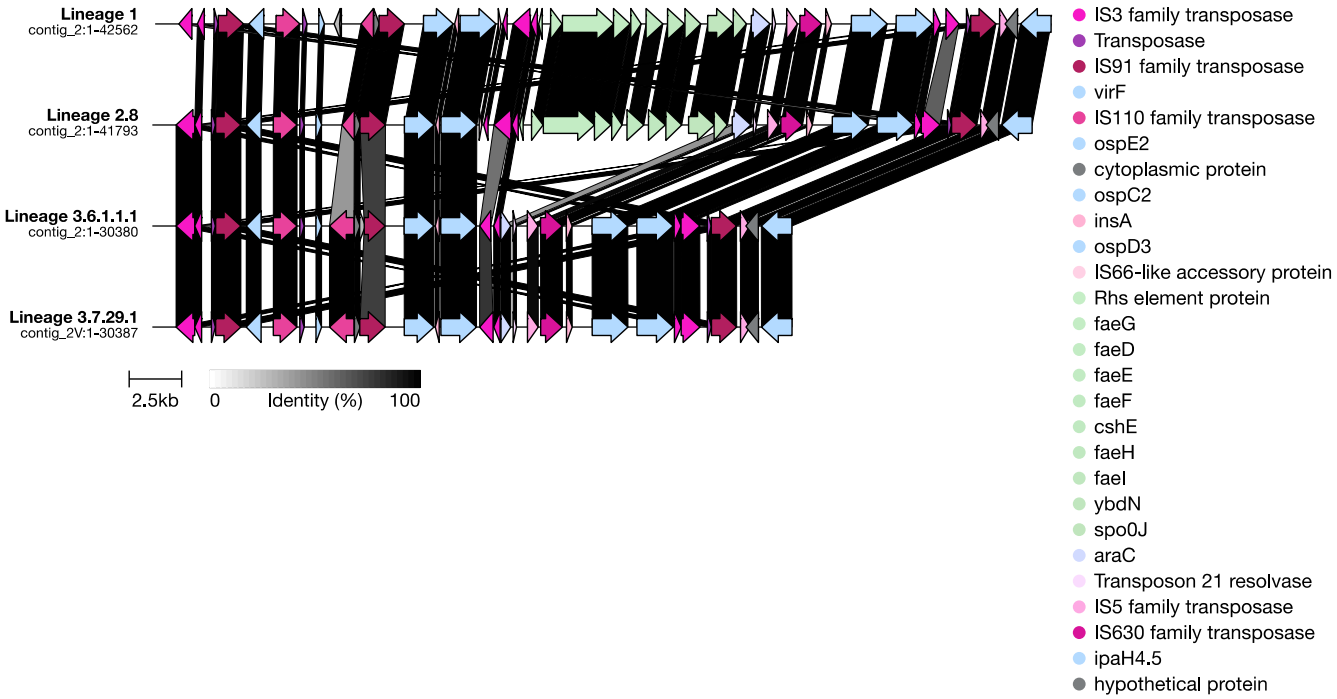


Figure 4.11. Clinker gene cluster comparison of the genomic region between IS3 and *ipaH4.5* depicting a deletion in representative lineage 3 genomes. Gene cluster comparisons were performed on representative genomes from lineage 1, 2, clade 3.6 and 3.7 using Clinker. Homologous regions are linked through the black/grey bars which also indicate the percentage identity. IS copies and related accessory genes are denoted in pink/purple, T3SS associated genes are coloured in blue and the deleted region is coloured in green.

Since the O-antigen and G4C have been previously implicated as key factors in the pathogenesis and survival of *S. sonnei*, the chromosomally encoded G4C cluster, and the pINV encoded O-antigen cluster were next compared. This revealed that the gene clusters encoding for both O-antigen and G4C were highly conserved between lineages, but the surrounding gene content varied in both cases. Upstream of the O-antigen gene cluster, an additional copy of an IS91 family transposase, and a gene annotated as encoding for a 'cytoplasmic protein' were identified in most clade 3.7 isolates, except for clade 3.7.16, which instead harboured an additional IS*630* copy (Fig 4.12). A hypothetical protein upstream of the O-antigen cluster was also identified in all lineage 2 and 3 genomes but absent from both

lineage 1 genomes. Similarly, a hypothetical protein upstream of the G4C gene cluster also was identified in all lineage 2 and lineage 3 genomes but absent from both lineage 1 genomes, but the gene cluster was otherwise conserved between all lineages (Fig 4.13).

Together, the data presented here highlight the presence of structural variation within *S. sonnei* lineages for the first time. In many cases, rearrangements were associated with flanking ISs which (in agreement with section 4.2.4) is suggestive of ongoing IS activity in lineages 2 and 3. I also show that genomic rearrangements are mostly concentrated to the chromosome in the form of deletions and inversions, whilst the structure of pINV was comparatively stable, except for the fixed loss of the *fae* gene cluster in lineage 3 plasmids.

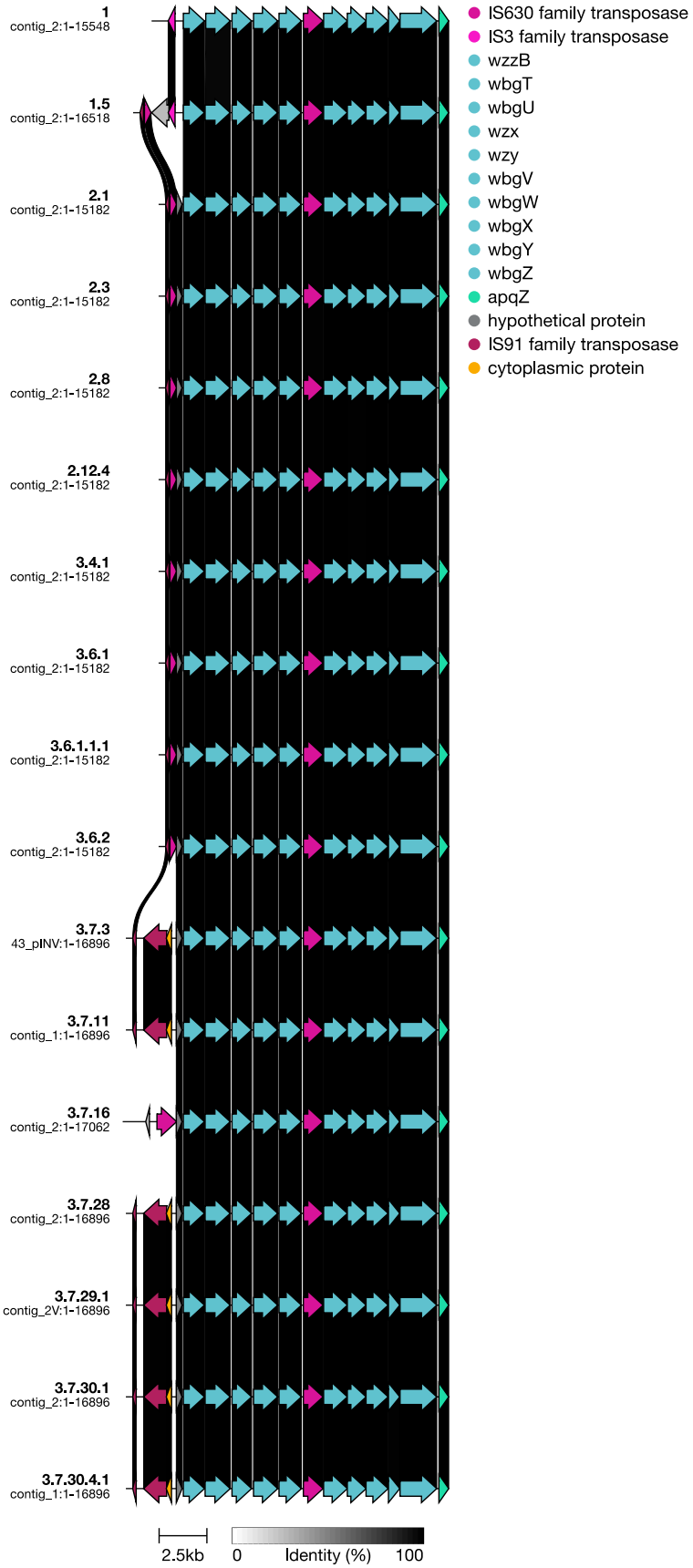


Figure 4.12. Clinker gene cluster comparison of the genomic region between IS91 and *apqZ* encoding for the O-antigen synthesis machinery.

Homologous regions are linked through the black bars which also indicate the percentage identity. IS copies and related accessory genes are denoted in pink/purple, hypothetical proteins are shown in grey and the O-antigen encoding gene cluster is coloured in blue. Numbers in bold denote the lineage to which the genome belongs.

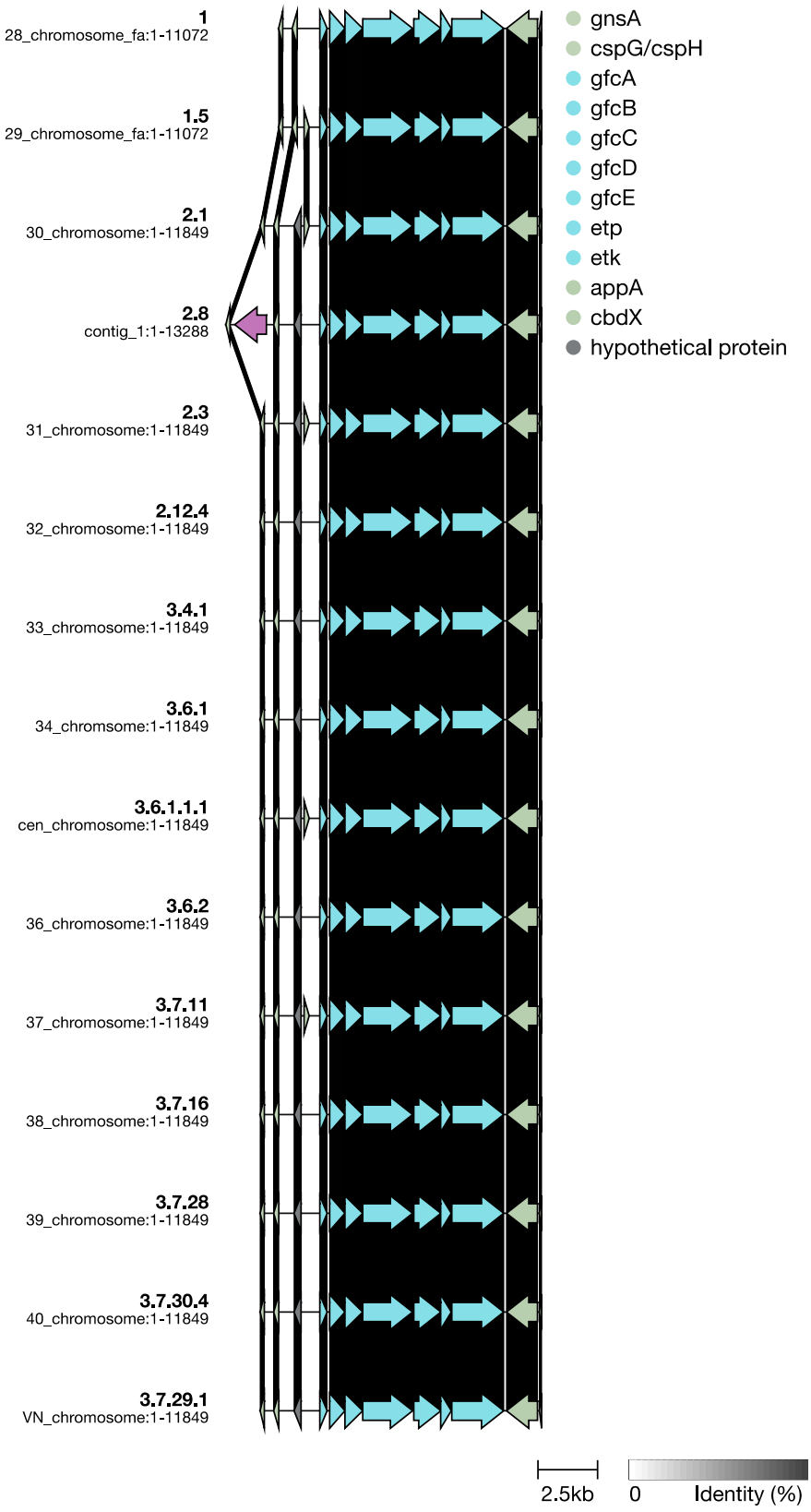


Figure 4.13. Clinker gene cluster comparison of the genomic region between *gnsA* and *cbdX* encoding for the group four capsule (G4C) synthesis machinery.

Homologous regions are linked through the black bars which also indicate the percentage identity. IS copies and related accessory genes are denoted in pink/purple, the G4C synthesis/assembly gene cluster is coloured in blue. Neighbouring genes are depicted in green and hypothetical proteins are depicted in grey. Numbers in bold denote the lineage to which the genome belongs.

4.3 Discussion

4.3.1. Overview

In this chapter, 16 new *S. sonnei* genomes (15 of which contain pINV sequences) were sequenced to completion and will be subsequently deposited into NCTC and RefSeq to act as a reference collection for epidemiologically relevant *S. sonnei* lineages. Completed genomes were characterised and compared based on genome content, which highlighted lineage dependent variations in genome size, the abundance and distribution of ISs and the number of pseudogenes. I next performed a whole genome alignment which illuminated variations in the genome structures of *S. sonnei* lineages for the first time. Overall, the results presented in this chapter provide novel insights into the genomic variation of *S. sonnei* lineages, providing evidence of continuing evolution and adaptation within *S. sonnei*.

4.3.2. Evidence for ongoing reductive evolution in *S. sonnei*

Reductive evolution has been key in shaping the transition from commensal *E. coli* towards pathogenic *Shigella* (Maurelli et al., 1998). Data presented here are supportive of ongoing genome reduction within *S. sonnei,* specifically within lineage 3. Overall, genome sizes and the number of predicted CDSs were smaller in lineage 3 isolates compared to lineages 1 and 2. Differences were, however, relatively small and differences were not always found to be statistically significant; future analysis would therefore benefit from the inclusion of additional lineage 1 and 2 representatives to highlight this difference with more clarity.

As well as smaller genome sizes, screening for virulence factors revealed that lineage 3 genomes also harboured fewer virulence genes overall. Differences were most evident in the number of adhesins and fimbrial subunits, with the loss of at least 10 genes fixed in all lineage 3 genomes. The presence of T3SS associated genes and iron acquisition genes were otherwise ubiquitous in pINV+ genomes, suggesting that they remain important for human infection. While this approach identifies well-described canonical virulence factors, it is important to acknowledge the limitations of using a database screening approach such as VFDB. These databases do not account for unknown or novel virulence factors and,

furthermore, do not consider the genomic context, which can include alterations in gene expression or regulatory pathways that may influence virulence.

The convergent loss of immunogenic features is common in many human-adapted pathogens (Weinert and Welch, 2017), and has been documented in the *Shigella* genus previously, with all subgroups having lost genes encoding for flagellin subunits (Yang et al., 2005). The conserved disruption of the *csgD* gene has previously been reported in other *Shigella* and EIEC subgroups either through deletions or insertions (Sakellaris et al., 2000), suggestive of a strong selective pressure against the production of curli in *Shigella*. Likewise, the inactivation of type 1 fimbriae (encoded by the *fim* gene cluster) has been reported to have occurred via at least three separate mechanisms, affecting all *Shigella* and some EIEC subgroups (Bravo et al., 2015). Current literature suggests that curli and type 1 fimbriae production is non-functional in all *Shigella* subgroups (Sakellaris et al., 2000), however, most studies have not explored intra-lineage variation, and none have included lineage 1 *S. sonnei* isolates. It would therefore be of interest to test their functionality in future, to see if lineage 1 represents the first *Shigella* subgroup described to have a functioning fimbrial apparatus. Interestingly, the K88 fimbriae (encoded by the *fae* gene cluster) has not been previously reported in other *Shigella* (although it is described as a key virulence factor in enterotoxigenic *E. coli* (ETEC) (Payne et al., 1993)). Its fixed loss in lineage 3 suggests that its presence is disadvantageous for *S. sonnei* and is supportive of an overall trend towards the loss of immunogenic features such as fimbriae which may contribute to the formation of pathovariants. The conserved loss of such features indicates that lineage 3 *S. sonnei* may be better at immune evasion, as compared to its lineage 1 and 2 counterparts, a fact that could contribute to its enhanced persistence and success.

The incidence of MDR (defined by resistance to three or more antibiotic classes) in *Shigella* is a predominant feature of currently circulating isolates (Baker and Scott, 2023), and in agreement with this, the presence of AMR determinants was widespread in all genomes analysed here. However, MDR was common in all lineages, and not necessarily restricted to

lineages considered epidemiologically successful. This therefore indicates that whilst AMR is a defining feature in the success of lineage 3 *S. sonnei*, it likely is not the only factor driving success. An association between colicin carriage and epidemiological success in *S. sonnei* was recently reported (Leung et al., 2024, De Silva et al., 2023) but this analysis did not include representatives from lineage 1 and 2. My findings highlighted the presence of a diverse range of colicins in all three lineages, which suggests that colicins are also not the only factor driving differences in epidemiological success. In future, experimental bacterial killing assays could be useful to identify any lineage-dependent differences in the scale of colicin-mediated killing, as this could differ.

Finally, pangenome analysis revealed limited variation in the number of HGCs between lineages, with the majority of genes being conserved across all lineages. There were few lineage specific HGCs, and most were annotated as either ISs or hypothetical proteins, a finding consistent with a previous analysis highlighting only minor gene content variation between *S. sonnei* lineages (Holt et al., 2012). This suggests that variation in *S. sonnei* success may not be driven by the binary presence or absence of a gene, but instead variations in gene functionality (as indicated by an increased number of both ISs and pseudogenes in lineage 3 *S. sonnei*) mediated by genome degradation, which would not necessarily be captured with a pangenome approach.

4.3.3. Variations in the burden of insertion sequences (ISs) in *S. sonnei* lineages

Bacteria which have emerged as highly specialised pathogens frequently have a much higher abundance of ISs (Moran and Plague, 2004). Prior to this study, the burden of ISs had only been explored in the chromosomal sequences (and not pINV) of *S. sonnei* lineages, and prior analysis was largely derived from short-read sequences, which can lead to the underestimation of the burden. Consistent with the hypothesis that lineage 3 *S. sonnei* is undergoing reductive evolution, I found that the chromosomal IS burden was greatest in lineage 3, whilst lineage 1 harboured the fewest. The contribution of individual IS families to the total IS burden in each lineage also varied, with the relative proportion of IS630 and IS3

increasing in lineage 3 genomes, which may be indicative of the proliferation of these IS families, consistent with previous results demonstrating the expansion of IS3 family elements in other *Shigella* subgroups (Hawkey et al., 2020, Seferbekova et al., 2021). In pINV, the lineage-dependent differences in IS abundance were smaller (owing to its much smaller size) and the opposite trend was observed, with lineage 1 harbouring the greatest burden of ISs. This result is surprising, especially given that lineage 1 pINVs are ~20,000 bp larger than lineage 1 and 2, which would indicate more IS activity and thus more deletion events in lineage 3. However, a case whereby several deletion events involving ISs have occurred, could be imagined for lineage 3, which would result in fewer IS copies overall. As discussed in section 4.3.2, the inclusion of more genomes from lineages 1 and 2 would be beneficial in future analysis to confirm whether differences in IS abundance are conserved. However, this would likely require the targeted sequencing and completion of these genomes since these genotypes are now infrequently detected in countries where the routine sequencing of *Shigella* occurs.

The proliferation of IS in *Shigella* is directly correlated with gene pseudogenisation (Yang et al., 2005) and consistent with this, I found that the total number of pseudogenes was highest in lineage 3, where IS abundance was also the greatest. Collectively, these results support the notion that IS mediated genome decay is ongoing in *S. sonnei*. An in-depth analysis of the genes undergoing pseudogenisation, and their biological functions, will be important in future to fully characterise the impact of genome decay on the fitness of respective *S. sonnei* lineages, and to understand how it is shaping their evolution. The accumulation of pseudogenes is a common signature of pathogens which have undergone specialisation (Moran and Plague, 2004) and likewise, it has previously been postulated that an increase in niche specificity results in accelerated genome decay in *Shigella* (Hershberg et al., 2007). In this case, given the decrease in genome sizes and increase in both IS and pseudogene abundance, it is tempting to speculate that lineage 3 *S. sonnei* is more specialised to its human host than lineages 1 and 2.

4.3.4. Variability in the structure of *S. sonnei* genomes

One of the most striking consequences of IS accumulation and expansion is the subsequent genomic instability that can manifest, a process that has been well documented in host-restricted *Bordetella* (Parkhill et al., 2003, Weigand et al., 2017, Park et al., 2012). Genomic rearrangements have previously been documented for *Shigella* as a whole (Seferbekova et al., 2021), but a comparison of structural variants in *S. sonnei* lineages has not previously been carried out, owing to a lack of completed genomes. Alignment of the newly completed genomes generated for this study revealed the presence of multiple structural rearrangements within *S. sonnei*, mostly in the form of deletions and inversions. In most cases, rearrangements were conserved across lineages (occurring in either one or two separate lineages) and could be linked to flanking IS sequences, suggestive of recombination. Consistent with my analysis of IS proportion (indicating a proliferation of IS3 in lineage 3) most structural rearrangements could also be linked to flanking IS3 family transposases; an IS family that has been associated with frequent genome inversions in other Enterobacteriaceae (AlKindy and Guyeux, 2022). A more detailed analysis of the sequence homology of flanking IS sequences will be needed to directly attribute the rearrangements documented here to IS mediated events, but nevertheless, the presence of flanking ISs is highly suggestive of IS mediated activity.

In some cases, rearrangements were documented in different lineages in similar genomic regions, but the actual content of the rearrangement differed. This could suggest that these rearrangements have occurred convergently and may be indicative of genomic hotspots where rearrangements are more likely to occur. It would be of interest to compare such regions to other *Shigella* genomes to determine if these rearrangements were also occurring in other genomes, or if this is a feature unique to *S. sonnei*.

Genomic rearrangements have been linked to driving changes in gene expression patterns, in cases whereby the promoter region is subsequently modified (Waters et al., 2022, Sousa et al., 1997), and where the IS harbours an outward directed promoter that can influence neighbouring gene expression (Chandler and Mahillon, 2007). In the case of *Neisseria*

meningitidis, the insertion of an IS*1301* copy into its capsule locus modified the expression of its capsule biosynthesis gene cluster, which subsequently altered its ability to resist bactericidal antibodies (Kugelberg et al., 2010). Similarly, in addition to mediating genomic rearrangements, the presence of ISs was also found to influence neighbouring gene expression in *B. pertussis* (Amman et al., 2018). Here, the alignment of key virulence factors, such as the O-antigen and G4C, revealed limited structural diversity in the encoding gene clusters but highlighted some lineage-dependent variations in the surrounding IS content. This provides evidence that IS activity is occurring within these key regions and it will therefore be of interest to test whether such rearrangements have any functional impact on gene expression in future. Finally, the lineage-specific nature of some genomic rearrangements reported here might be indicative of differential niche occupation or selection pressures, since rearrangements are known to contribute to bacterial adaptability.

4.3.4. Conclusions

In conclusion, data presented here support the hypothesis of ongoing IS mediated genome reduction in the evolution of *S. sonnei*; my results demonstrated smaller genomes, a greater number of IS pseudogenes, and a greater frequency of structural rearrangements in the completed genomes of epidemiologically important lineage 3 isolates. The functional characterisation of lineage dependent variations identified here will be important in future to determine whether lineage specific genome variations also translate to functional differences that may confer a fitness advantage.

Chapter 5. Experimental characterisation reveals increased virulence and an increased stress tolerance in lineage 3 *Shigella sonnei*

5.1. Introduction

5.1.1. *S. sonnei* infection models

S. flexneri has been the preferred model for *Shigella* infection for many decades, due to its more stable pINV and thus enhanced reliability as a lab workhorse. Where *S. sonnei* has been used, infection models have mostly been restricted to lab-adapted isolate 53G (of genotype 2.8), which is no longer reflective of the current circulating *S. sonnei* genotypes. The current understanding of *S. sonnei* virulence and pathogenesis has therefore been largely extrapolated from *S. flexneri* infection, but recent studies have highlighted some key differences between the two species. *S. flexneri* is well-known for its inefficiency at infecting static monolayers of cells (Carayol and Tran Van Nhieu, 2013) while *S. sonnei* is regarded as even less efficient (Niesel et al., 1985). This inefficiency, coupled with the high rates of pINV instability, has hindered the development and use of *S. sonnei* infection models. Considering this, along with a shift towards *S. sonnei* dominance and the upsurge in AMR infections, there is a clear need for the further development and use of *S. sonnei* infection models to fully understand its pathogenesis and how this might contribute to its recent dominance over *S. flexneri*.

Studies infecting macrophages *ex vivo* have revealed that *S. sonnei* is less likely to be phagocytosed and as a result, causes less pyroptosis and subsequent inflammation than *S. flexneri;* such differences were found to be dependent on its O-antigen (Watson et al., 2018). In a similar manner, the O-antigen of *S. sonnei* was found to be important in resisting phagolysosome acidification and neutrophil cell death in a zebrafish infection model (Torraca et al., 2019). The *S. sonnei*-zebrafish model has been highly valuable in teasing out differences between *Shigella* species, highlighting that *S. sonnei* is more virulent than

S. flexneri in zebrafish, and that it is also more likely to establish long-term, persistent infections (consistent with what is observed in humans) (Torraca et al., 2023). Overall, the *S. sonnei*-zebrafish model is well placed to explore differences in *S. sonnei* lineages and could contribute to unveiling drivers of epidemiological success.

5.1.2. Aims

This chapter investigates functional variations in a collection of epidemiologically relevant *S. sonnei* isolates (established and sequenced in Chapter 4). The overall aim was to identify any lineage-dependent variations that contribute towards the domination of lineage 3 in the *S. sonnei* epidemiological landscape. The specific aims of this chapter were: to characterise variations in the pathogenicity of *S. sonnei* lineages and to experimentally test underlying factors that may contribute to differences in virulence and epidemiological success.

5.2. Results

5.2.1. Lineage 3 *S. sonnei* is most virulent in a zebrafish infection model

To assess the virulence of *S. sonnei* genotypes sequenced in Chapter 4, four representative isolates were first selected to test in the zebrafish infection model. The four isolates comprised of lineage 1.5, the well characterised lineage 2.8 (53G) and representatives from the epidemiologically successful clades 3.6 and 3.7 (3.6.1.1.1 and 3.7.29.1). Zebrafish larvae at 3 dpf were infected with 2,000 CFU of *S. sonnei* in the HBV and incubated at the standard incubation temperature of 28.5 °C, or the increased temperature of 32.5 °C, considering the thermoregulated virulence I demonstrated for ST99 EIEC (see Chapter 3 (Miles et al., 2023)).

At 28.5 °C, there were no significant differences in larvae survival by 48 hpi, with ~65-70% survival observed for all lineages (Fig 5.1 A). However, at 32.5 °C, the survival of larvae infected with lineage 3.6 or 3.7 isolates reduced to ~30%, significantly less than larvae infected with lineage 1.5 or 2.8, which exhibited ~50 and 60% survival respectively (Fig 5.1 B). To determine whether lineage 3 isolates had a replicative advantage, bacterial burden was next enumerated from infected larvae at 6 and 24 hpi. A greater bacterial burden for all lineages was measured at 32.5 °C, as compared to 28.5 °C, but differences were not lineage dependent at either time-point or temperature (Fig 5.1. C,D). Since differences between lineages were only observed at the higher temperature, all further infections were performed at 32.5 °C.

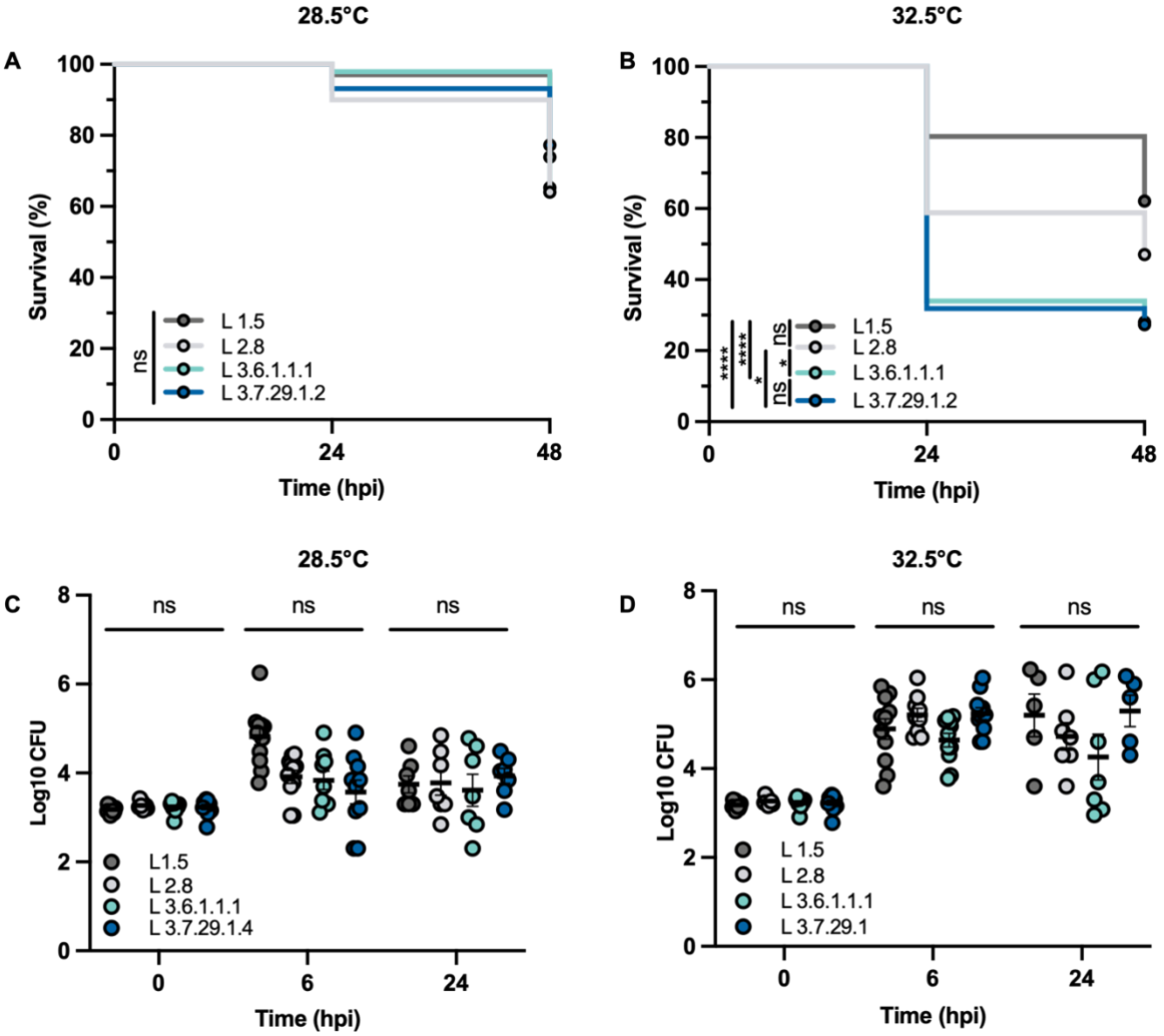


Figure 5.1. *S. sonnei* infections of zebrafish larvae at 28.5 °C and 32.5 °C. A, B) Survival curves of zebrafish infected with 2000 CFU of different *S. sonnei* lineages. Larvae were incubated at 28.5 °C and 32.5 °C respectively, and survival was measured at 24 and 48 hours post infection (hpi). Data is pooled from three independent experiments (n ≥ 15 larvae per condition/experiment). Significance was determined using the log-rank Mantel-Cox test, *p<0.0332; ****p<0.0001. **C,D)** CFU counts measured at 6 and 24 hpi for infected larvae incubated at 28.5 and 32.5 °C respectively. Data is pooled from three independent experiments (n=3 larvae per timepoint in cases where larvae remained viable). Significance was tested using a two-way ANOVA with Tukey's correction applied. Error bars represent the mean ± SEM. L = lineage, ns = non-significant.

To determine if lineage-dependent differences in virulence were conserved in additional isolates of the same lineage, infections were next repeated to include more representatives from each lineage, with lineage 2.8 included in all experiments for comparative purposes. Neither the lineage 1 isolates nor the additional lineage 2 isolate (2.1) showed any difference in virulence when compared to lineage 2.8 (Fig 5.2 A,B). More variation was observed in lineage 3, with clade 3.4 (~40% survival) being less virulent than clade 3.6 (~20% survival) (Fig 5.2 C). However, virulence between all three tested subclades of 3.6 remained consistent, and significantly more virulent than the comparative strain, lineage 2.8. Intra-clade variation was detectable for clade 3.7, with infected larvae exhibiting ~15-30% survival. In all cases, however, there were significant differences in survival when compared to comparative strain, lineage 2.8 (Fig 5.2 D).

These results highlight the sensitivity of the zebrafish infection model in comparing the pathogenicity of closely related lineages. Moreover, epidemiologically successful clades 3.6 and 3.7 were, in all cases, found to be more virulent than representatives from lineages 1 or 2 and even other lineage 3 subclades. Since there were no significant differences between lineages 1 and 2 and limited intra-clade variation in lineage 3, further experiments are performed with representative lineages 2.8, 3.6.1.1.1 and 3.7.29.1.2 unless indicated otherwise.

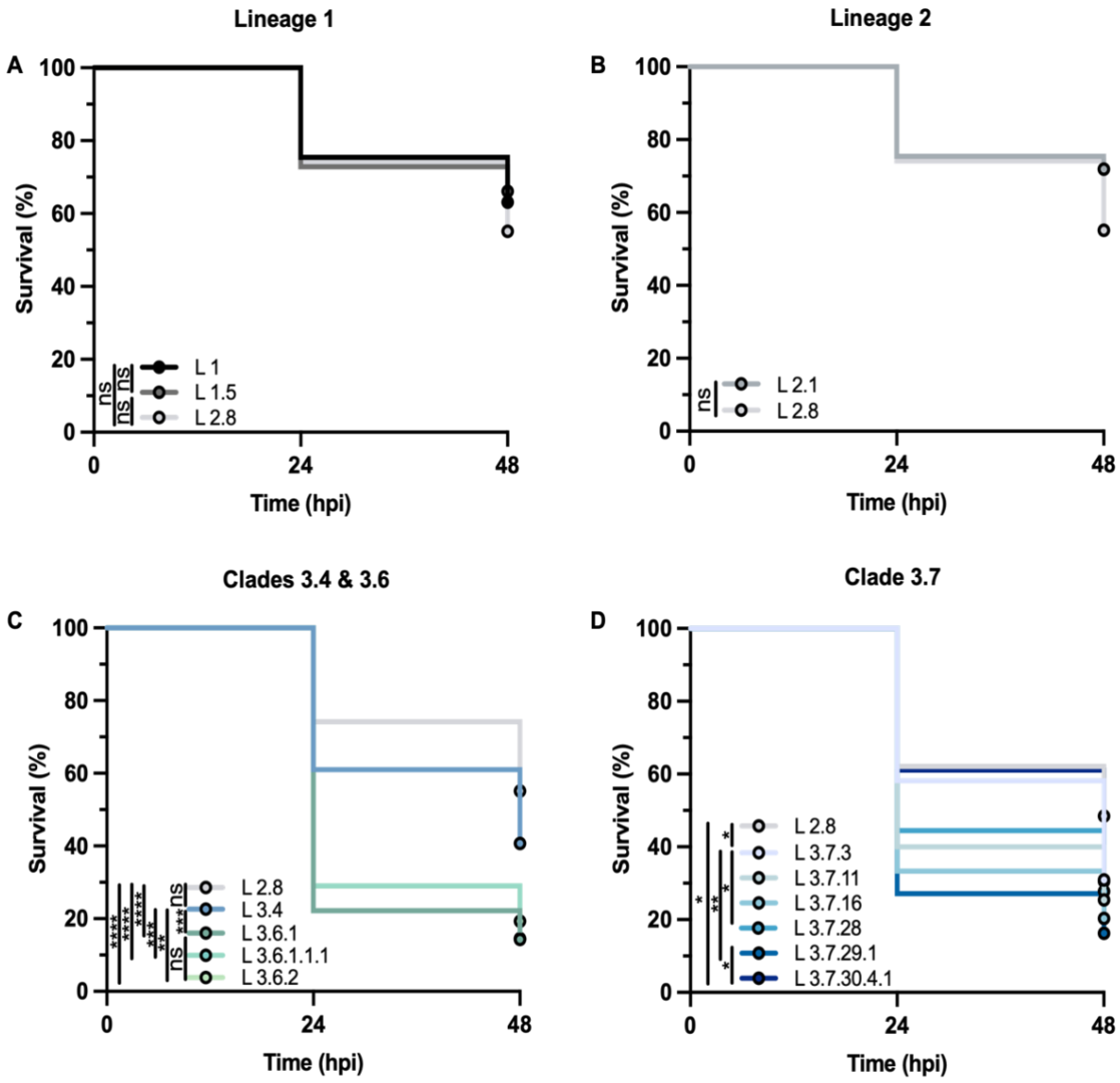


Figure 5.2. *S. sonnei* infections of zebrafish larvae including additional representatives from each lineage. Larvae were infected with 2000 CFU of different *S. sonnei* lineages, incubated at 32.5 °C and survival was monitored at 24 and 48 hours post infection (hp)i. **A)** Survival curve of larvae infected with lineage 1 representatives. **B)** Survival curve of larvae infected with lineage 2 representatives. **C)** Survival curve of larvae infected with representatives from clades 3.4 and 3.6. **D)** Survival curve of larvae infected with representatives from clade 3.7. Lineage 2.8 was included in all experiments for comparative purposes. Not all p values could be displayed for clade 3.7 due to space limitations, but all L 3.7 isolates were significantly more virulent than L 2.8 Data is pooled from three independent experiments (n ≥15 larvae per condition/experiment). Significance was determined using the log-rank Mantel-Cox test, *p<0.0332; **p<0.0021; ***p<0.0002; ****p<0.0001. L = lineage, ns = non-significant.

Recent studies using the zebrafish infection model demonstrated that *S. sonnei* 53G (genotype 2.8, lab strain) disseminates from the HBV down the neural tube at a much higher frequency than *S. flexneri* (Lensen et al., 2023), while a prior rabbit infection model showed similar dissemination from the infection site in a G4C-dependent manner (Caboni et al., 2015). To investigate if this phenotype holds true for clinical *S. sonnei* isolates, infections were performed with fluorescently labelled bacteria, larvae were imaged and then assessed for dissemination events. Dissemination, defined as the spreading from the HBV into the neural tube, was observed for all three isolates tested (Fig 5.3), and consistent with data showing increased virulence, higher dissemination levels were recorded for clades 3.6 and 3.7, although only clade 3.7 was found to be statistically significantly different from lineage 2.8.

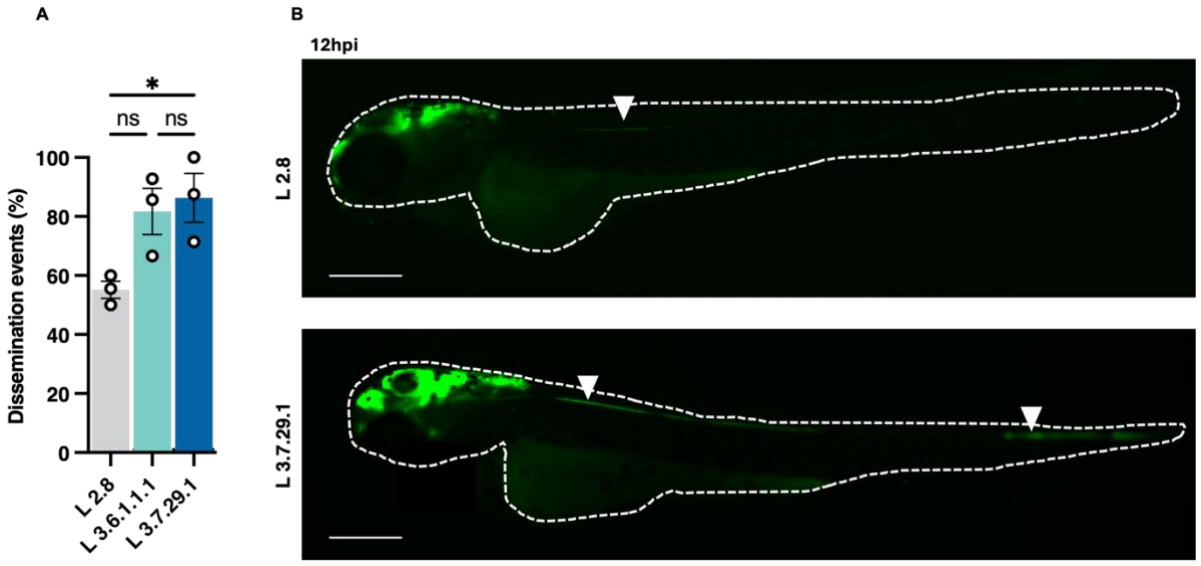


Figure 5.3. Dissemination of *S. sonnei* from the HBV infection site in infected zebrafish larvae. Larvae were infected with GFP labelled bacteria and dissemination events were recorded as spreading from the hindbrain ventricle (HBV), regardless of the extent of bacterial spread. **A)** Percentage of dissemination events recorded at 12 hours post infection (hpi) from three independent experiments (where $n \geq 8$ larvae per group). Data is presented as the mean ± SEM and significance was tested using a one-way ANOVA with Tukey's correction applied, *p<0.0332, ns = non-significant. **B)** Representative images depicting bacterial spread along the neural tube by L 2.8 (top) and L 3.7.29.1), dissemination events are indicated with a white arrow. White dashed line indicates the zebrafish outline. Images represent a single Z slice taken on a Leica M205FA stereomicroscope at X0.3 magnification. Scale bar = 500 µm.

Considering the instability of pINV in *S. sonnei* under laboratory conditions, I hypothesised that pINV in clades 3.6 and 3.7 is more stable than in other lineages, explaining the observed increase in virulence. Therefore, I cultured various *S. sonnei* isolates, either grown in broth or passaged through zebrafish larvae, and calculated the ratio of Congo red-positive (CR+; indicating pINV+), to the total number of colonies. There were no lineage dependent differences in pINV stability detectable if the bacteria were grown *in vitro* (Fig 5.4 A). Compared to *in vitro*, pINV stability *in vivo* increased by ~40% for all isolates tested and reduced pINV stability was detectable in clades 3.6 and 3.7 (compared to lineage 2.8). Overall, this shows that increased virulence in zebrafish cannot be linked to a more stable pINV for clades 3.6 and 3.7. Furthermore, these data prove that pINV is more stable in our zebrafish infection model than during *in vitro* culturing, supporting pINV as a key driver of *S. sonnei* virulence *in vivo*.

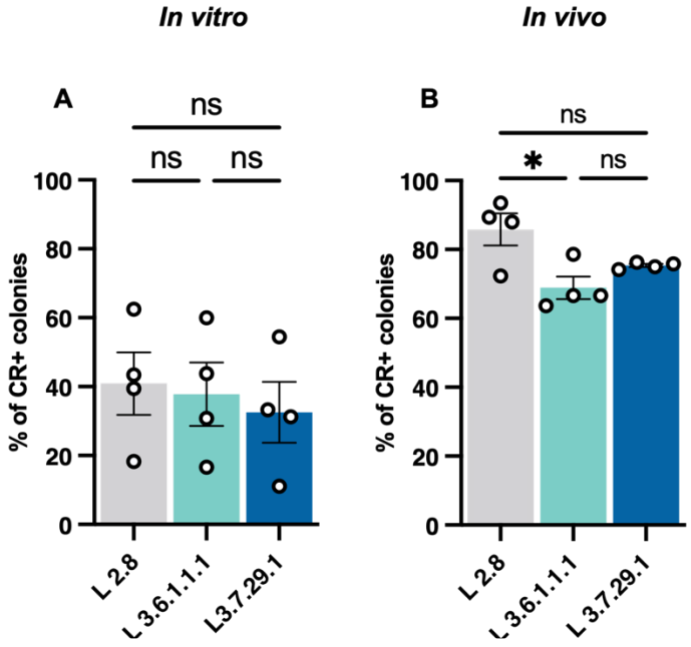


Figure 5.4. Stability of pINV *in vitro* and *in vivo*. pINV stability was measured by taking a ratio of Congo Red (CR)+ (and thus pINV+) colonies to the total number of colonies. **A)** For measuring stability *in vitro*, bacteria were grown in liquid broth and plated directly onto CR supplemented TSA. **B)** For measuring stability *in vivo*, a homogenate of infected larvae was plated onto CR supplemented TSA. Data are representative of four independent experiments, with technical triplicates for each measurement also performed. Data is representative of the mean ± SEM and significance was tested using a one-way ANOVA with Tukey's correction applied, *p<0.0332, ns = non-significant.

Given the importance of the T3SS in *Shigella* virulence (Schroeder and Hilbi, 2008), and my results indicating IS activity within the T3SS encoding region (presented in section 4.2.5), I hypothesised that variations in T3SS gene expression or effector secretion contribute to the virulence of lineage 3. Analysis of gene expression revealed no differences in the expression of master regulator genes *virF* and *virB*, or of T3SS effector *ipaB* (Fig 5.5 A,B,C). Likewise, the analysis of protein secretion from bacteria grown *in vitro* revealed no overall change in the secretion of effector proteins between the *S. sonnei* lineages (Fig 5.5 D). Together, these results suggest that T3SS activity does not drive the observed differences in virulence in *S. sonnei*.

Overall, these results are indicative of increased virulence in epidemiologically successful lineage 3 *S. sonnei*, with increased bacterial dissemination and reduced zebrafish survival. Bacterial burden, pINV stability and T3SS activity were all investigated as potential drivers of increased virulence, but neither resulted in any lineage-dependent distinctions. Considering this, I next focussed on the impact of zebrafish host-response to infection with different *S. sonnei* lineages.

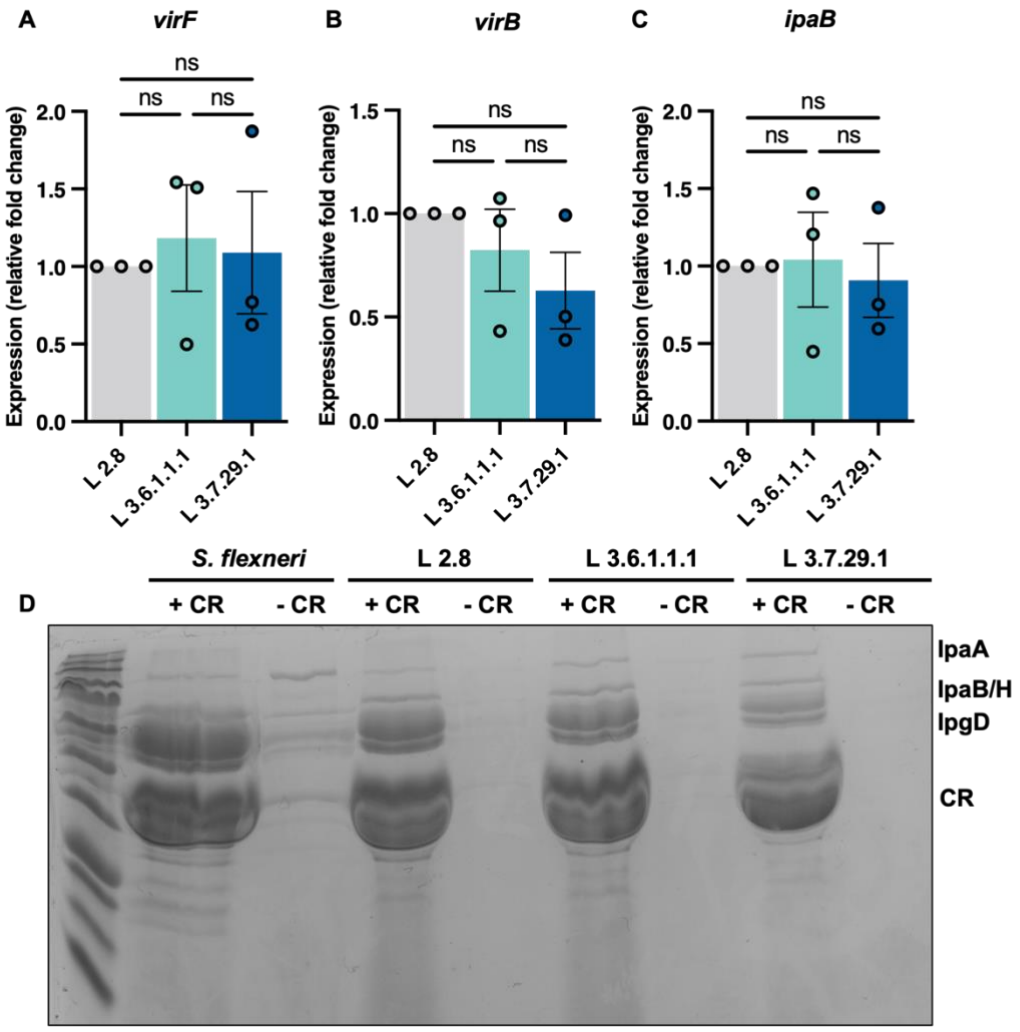


Figure 5.5. Expression and secretion of the type three secretion system (T3SS). A,B,C) Relative mRNA expression of T3SS genes *virF, virB* and *ipaB in vitro.* Bacteria were grown at 37°C in 5 mL TSB, with shaking at 400 rotations per minute until mid-exponential phase (OD 0.5-0.6), where an amount corresponding to $1x10^9$ cells was pelleted prior to RNA isolation. RNA was extracted from pelleted bacteria and qRT-PCR was performed on the resulting cDNA, relative gene expression was determined by normalisation to *rrsA* as a housekeeping gene, and then to L 2.8 using the delta delta Ct method. **D)** SDS-PAGE gel stained with Coomassie blue, showing secreted factors in the presence or absence of Congo red (CR). *S. flexneri* M90T was included as a positive control. qPCR data are representative of 3 independent experiments, with technical duplicates for each measurement also

5.2.2. Lineage 3 *S. sonnei* induces a stronger pro-inflammatory immune response

Genomic analysis indicated that lineage 3 isolates harbour fewer genes encoding for putative immunogenic components (i.e. those that have previously been described to elicit an immune response during infection) (section 4.2.2). Therefore, I tested if *S. sonnei* lineages were inducing differential immune responses during infection. Since *Shigella* is known to kill both neutrophils and macrophages (Schnupf and Sansonetti, 2019, Brokatzky et al., 2024), the total abundance of these immune cells in infected larvae was measured over time. In line with bacterial induced killing, a steady decrease in neutrophils per larva was observed between 3-12 hpi for all lineages (Fig 5.6 A), but lineage-dependent differences in the total number of neutrophils were not observed. Likewise, macrophages also decreased throughout the course of infection, but on a much smaller scale when compared to neutrophil depletion (and in most cases did not significantly differ from the uninfected control), consistent with previous findings indicating limited macrophage killing by *S. sonnei* (Watson et al., 2019, Torraca et al., 2023). As with neutrophils, there were no lineage-dependent changes in macrophage numbers over time (Fig 5.6 B), which suggests that differences in the capacity to cause cell death at the whole animal level was not the primary driver of virulence differences between lineage 2 and 3 *S. sonnei*.

The zebrafish HBV typically has very few residing neutrophils or macrophages (Moussouni et al., 2021), so can be used to measure the recruitment of immune cells to the infection site. I found that neutrophils and macrophages were recruited to the HBV of infected larvae as early as 3 hpi (Fig 5.6 CD), and neutrophils were recruited in greater abundance than macrophages. The number of macrophages recruited over time was similar for all isolates tested and did not differ significantly in larvae infected with lineage 2 or lineage 3 isolates. The number of neutrophils recruited rose steadily until 6 hpi for all lineages, but at both 3 and 6 hpi, more neutrophils were recruited by 3 hpi for clade 3.7, and by 6 hpi, more neutrophils were recruited for both lineage 3 isolates, as compared to lineage 2. Overall, these data highlight the dominant role of neutrophils in responding to *S. sonnei* infection of the HBV and significantly,

shows that neutrophils respond faster and in greater abundance to lineage 3 infections as compared to lineage 2.

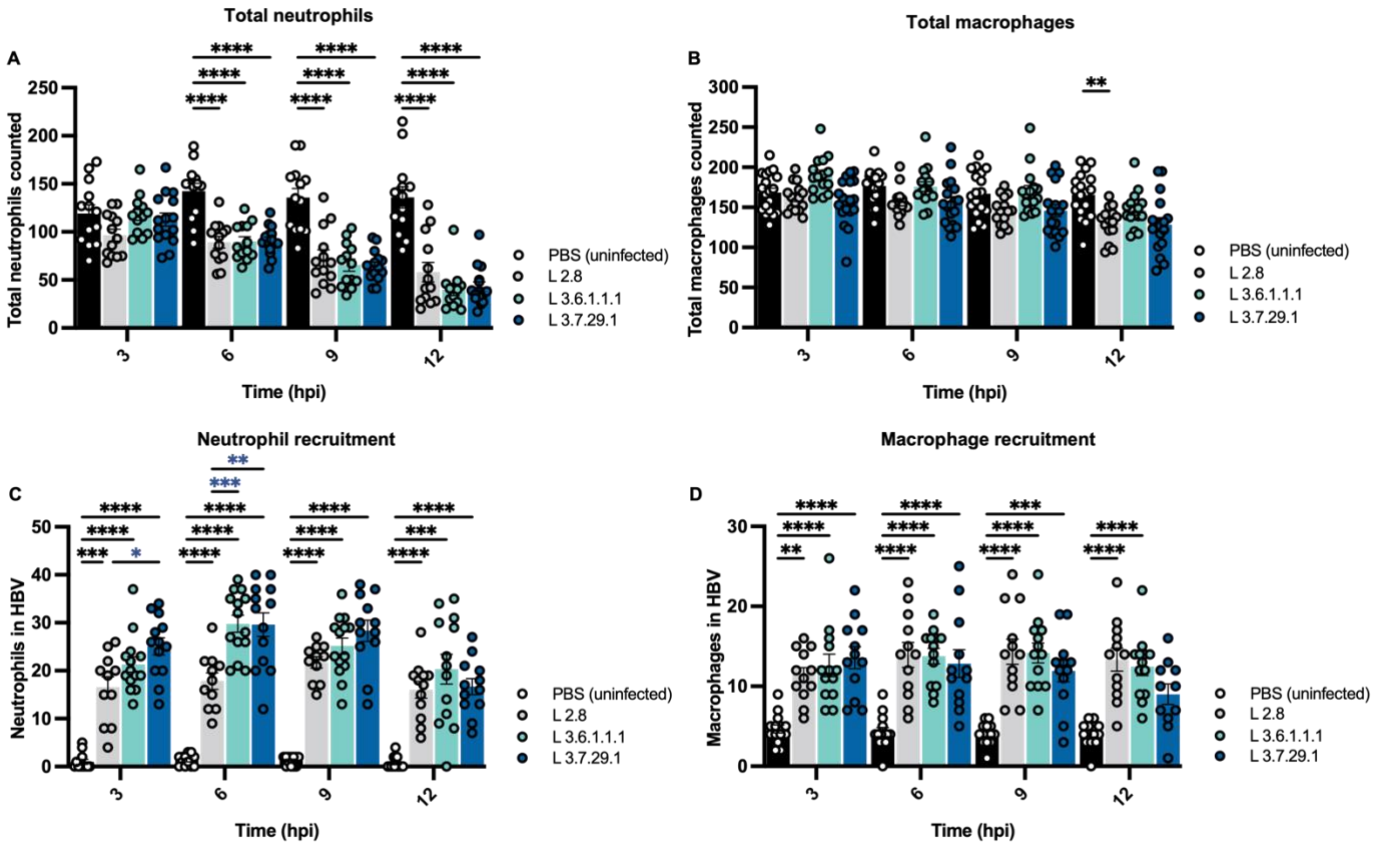


Figure 5.6. Dynamics of neutrophils and macrophages in *S. sonnei* infected zebrafish larvae. Zebrafish larvae with fluorescently tagged neutrophil or macrophages were injected with either *S. sonnei* or PBS as a control and images were taken every 3 hours post infection (hpi) from 3-12 hpi. Immune cells were counted from images at respective time points. **A)** Larvae were positioned on their lateral side and neutrophils in the whole larva were counted. **B)** Larvae were positioned on their lateral side and macrophages in the whole larva were counted. **C)** Larvae were positioned yolk-sack down, head-up and neutrophils recruited to the hindbrain ventricle (HBV) were counted. **D)** Larvae were positioned yolk-sack down, head-up and macrophages recruited to the HBV were counted. Images were taken on a Leica M205FA stereomicroscope. Data represents three independent experiments (with $n \geq 4$ individuals per group, per time point) and is presented as the mean ± SEM. Significance was tested with a two-way ANOVA with Tukey's correction applied. Only significance where $p \geq 0.05$ are displayed due to space limitations, *p<0.0332; **p<0.0021; ***p<0.0002; ****p<0.0001, asterisks shown in blue highlight lineage-dependent differences.

To link these findings with the global inflammatory response to infection, the expression of inflammatory cytokines was quantified at 6 hpi (where differences in neutrophil recruitment were the greatest). This revealed that expression of *cxcl8a*, a zebrafish homolog of interleukin-8 which plays a role in neutrophil chemotaxis (Oehlers et al., 2010), was found to be upregulated in all infected larvae and the expression levels were found to be significantly greater in lineage 3 infected larvae compared to lineage 2.8 (Fig 5.7 A). Similarly, *cxcl18b*, another (neutrophil chemokine (Torraca et al., 2017)), was upregulated (although not statistically significant) by ~75 fold in lineage 3 infected larvae (Fig 5.7 B). The expression of *il1b*, which has been linked to macrophage derived inflammation (Hasegawa et al., 2017), was generally lower than the other pro-inflammatory cytokines and no lineage dependent differences in expression were observed (Fig 5.7 C), consistent with a limited role for macrophages in *S. sonnei* infection control. Likewise, the expression of anti-inflammatory cytokines *il10a* and *il4a* was low and there were no lineage-dependent differences identified (Fig 5.7 D,E), further consolidating a role for neutrophil-mediated inflammation in the enhanced virulence of lineage 3 *S. sonnei*.

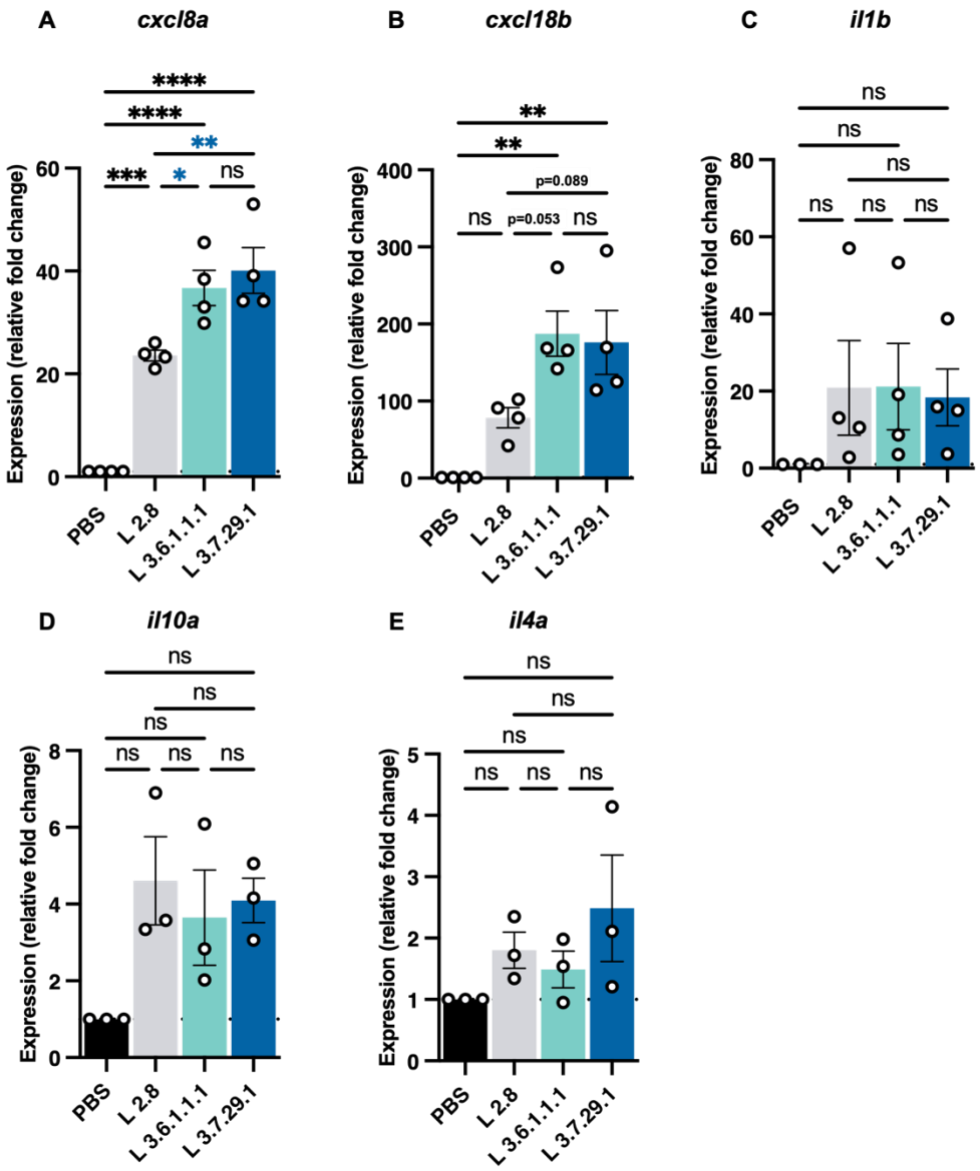


Figure 5.7. Relative mRNA expression of immune-related cytokines in infected zebrafish larvae. Zebrafish larvae were injected with either a PBS control or different *S. sonnei* lineages and RNA was then extracted from 10-15 individuals per group at 6 hpi. RT-qPCR was performed on the resulting cDNA and gene expression relative to uninfected samples was determined using the delta delta Ct method, with *eef1a1a* used as a housekeeping gene. **A,B,C)** The relative expression of pro-inflammatory cytokines *cxcl8a, cxcl18b* and *il1b.* **D,E)** The relative expression of anti-inflammatory cytokines *il10a* and *il4a.* Data represents at least three independent experiments and is presented as the mean ± SEM. Significance was tested with a one-way ANOVA with Tukey's correction applied, *p<0.0332; **p<0.0021; ***p<0.0002; ****p<0.0001, asterisks shown in blue highlight lineage-dependent differences. ns = non-significant, L = lineage.

To further investigate the contribution of inflammation to the enhanced virulence of lineage 3 *S. sonnei*, I tested whether chemical suppression of the host inflammatory response, using dexamethasone, could eliminate differences in virulence. Dexamethasone is a broad-spectrum anti-inflammatory drug which has previously been shown to reduce neutrophil recruitment in infected zebrafish larvae (Virgo et al., 2024). When treated with dexamethasone, the survival of all infected larvae was rescued (Fig 5.8 A) and the differences in survival between lineage 2 and lineage 3 isolates were reduced, but not eliminated. Bacterial burden determination by 6 hpi (Fig 5.8 B), proved that the suppression of inflammation does not affect bacterial replication, regardless of the lineage.

Collectively, these data highlight the role of inflammation in *S. sonnei* virulence, showing that zebrafish larvae survival can be rescued when inflammation is inhibited, and additionally supports a role for inflammation in the enhanced virulence of lineage 3. However, the inflammatory response cannot be attributed as the sole contributing factor, since lineage-dependency was not completely eradicated.

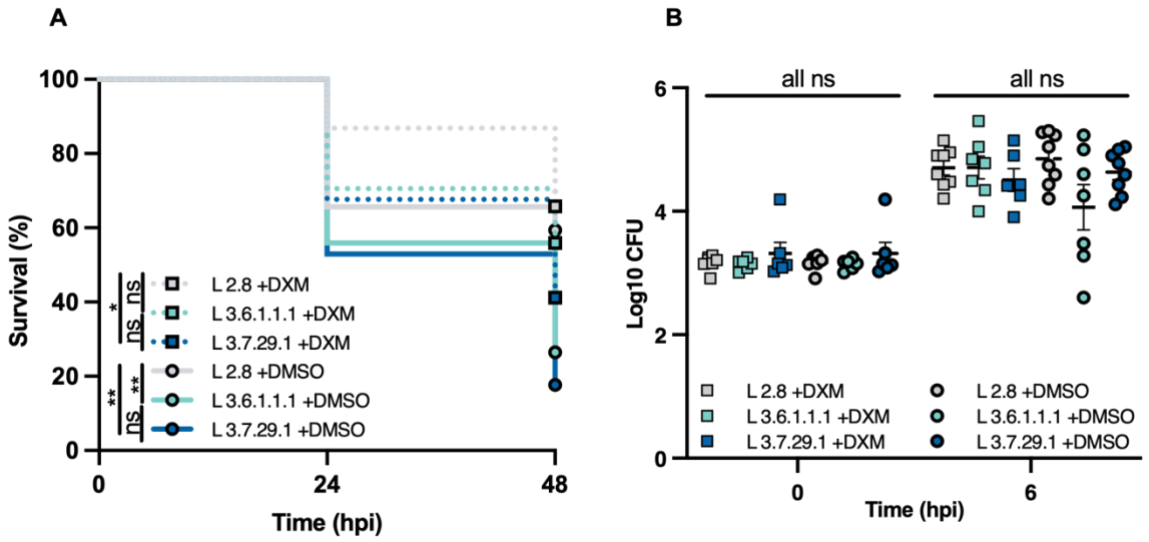


Figure 5.8. Chemical suppression of inflammation rescues the survival of *S. sonnei* infected zebrafish larvae. Zebrafish larvae were infected and bathed in either 50 µg/mL dexamethasone (DXM) (a chemical suppressant of inflammation), or an equal concentration of DMSO as a negative control. **A)** Survival of infected zebrafish bathed in dexamethasone or DMSO, survival was measured at 24 and 48 hpi. Data are representative of 3 independent experiments (with n ≥ 15 individuals per group) and significance was determined using the log-rank Mantel-Cox test. **B)** CFUs were enumerated at 0 and 6 hpi by the mechanical disruption of larvae, and plating of homogenate. Data is representative of two independent experiments (with n ≥ 4 individuals per group). All data is presented as the mean ± SEM. Significance was tested with a two-way ANOVA with Tukey's correction applied, *p<0.0332; **p<0.0021. ns = non-significant, L = lineage.

5.2.3. Lineage 3 *S. sonnei* are more tolerant of neutrophil killing *ex vivo*

Having observed variations in pathogenicity between *S. sonnei* isolates in zebrafish, it was next of interest to determine if differences could also be reiterated in human cells, given the nature of *Shigella* as a human adapted pathogen. In a first step, a gentamicin protection assay was used to test for distinctions in bacterial invasion and replication in human epithelial cells. Overall, rates of bacterial invasion into HeLa cells were low, consistent with what has previously been reported for *S. sonnei* (Watson et al., 2019). Time points of 1 hour 40 and 3 hours 40 were chosen to represent invasion and replication respectively. 1 hour 40 reflects 40 minutes of protocol time (10 minutes of centrifugation, 30 minutes of incubation) and 1 hour

of gentamicin treatment, the standard treatment time recognised to kill extracellular bacteria; 3 hours 40 minutes includes 40 minutes of protocol time with a longer gentamicin treatment of 3 hours to allow for sufficient intracellular replication (Small et al,. 1987). At 1 hour 40, there were no statistically significant differences in bacterial invasion between lineage 2 and lineage 3 isolates, and if anything, lineage 3 isolates appeared to be less invasive than lineage 2 (Fig 5.9 A). Similarly, at 3 hours 40 no variations in bacterial replication were identified between *S. sonnei* lineages (Fig 5.9 B,C), a finding that is consistent with bacterial burden data from zebrafish infections. Overall, these data show that lineage 3 does not invade HeLa cells more efficiently, and does not have a replicative advantage once intracellular, suggesting that interactions with other cell types may be driving lineage-dependent virulence.

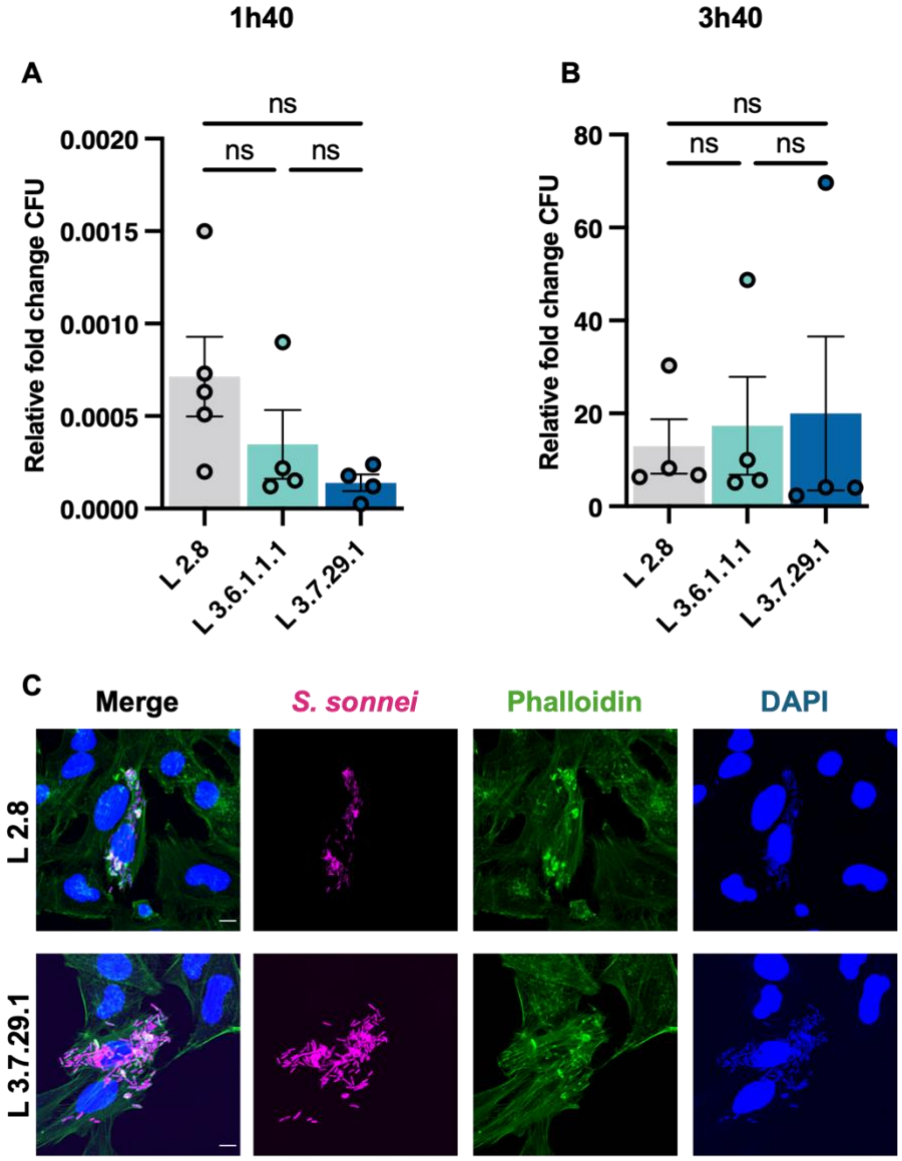


Figure 5.9. Invasion and replication of *S. sonnei* lineages in HeLa cells. HeLa cells were infected with *S. sonnei* at a multiplicity of infection (MOI) of 100. Cells were centrifuged at 500 x *g* for 10 minutes following inoculation and infected for 30 minutes before treatment with gentamicin. **A)** Cells were lysed, and bacteria were recovered at 1 hour 40 to determine invasion. The total number of bacteria recovered was normalised to the inoculum to determine relative fold change. **B)** Bacteria were recovered from lysed cells at 3 hours 40 and normalised to the total number of bacteria at 1 hour 40 to determine bacterial replication relative to the number of bacteria that invaded. Data are representative of four independent experiments and were performed in technical duplicates, error bars represent the mean ± SEM. Significance was tested with a one-way ANOVA with Tukey's correction applied. ns = non-significant, L = lineage. **C)** Representative airyscan confocal image of *S. sonnei* infected HeLa cells at 3 hours 40. *S. sonnei* is shown in magenta, phalloidin (F-actin) is shown in green and DAPI is shown in blue. Scale bar = 2 µm.

Neutrophils have been reported to have a primary role in *S. flexneri* control (Arena et al., 2015) and considering results highlighting lineage-specific neutrophil recruitment in zebrafish, I next infected primary human neutrophils with *S. sonnei* and tested bacterial survival. The total number of bacteria at 1 hour (relative to the inoculum) decreased in all cases, suggestive of bacterial killing by neutrophils (Fig 5.10 A,B). Strikingly, lineage 3 isolates appeared to be more resistant to neutrophil-mediated killing, with ~4-fold more bacteria recovered as compared to lineage 2.8. To investigate if lineage 3 was more cytotoxic to neutrophils, a lactate dehydrogenase (LDH, an enzyme that is released upon cell lysis) assay was performed to quantify the extent of neutrophil killing. For all lineages, ~20% LDH release (when compared to a fully lysed control sample) was observed and there were no lineage-dependent differences in LDH release (Fig 5.10 C), consistent with zebrafish results indicating no differences in immune cell killing (see section 5.2.2). Samples were also taken for qRT-PCR to quantify cytokine expression in infected human neutrophils. Here, minimal expression for all cytokines tested (*il8, il1b, tnfa* and *il10*) and no lineage-dependency in cytokine expression was observed (Fig 5.10 D). An enzyme linked immunosorbent assay (ELISA) for IL-1β was performed on the supernatant of infected neutrophils to measure cytokine release, and in line with neutrophil qRT-PCR results, no release was detected from either infected or uninfected cells (data not shown). This data contrasts cytokine expression results from zebrafish, where a significant upregulation in *cxcl8a* was observed, but this could be due to the time of 1 hpi being simply too early to detect a significant upregulation in cytokine expression or could suggest that the *cxcl8a* expression in zebrafish is epithelial and not neutrophil derived.

Overall, these *in cellulo* data using isolated human neutrophils agree with results obtained from zebrafish infection, indicating that lineage 3 shows altered interactions with neutrophils and increased tolerance to neutrophil-mediated killing compared to lineage 2.

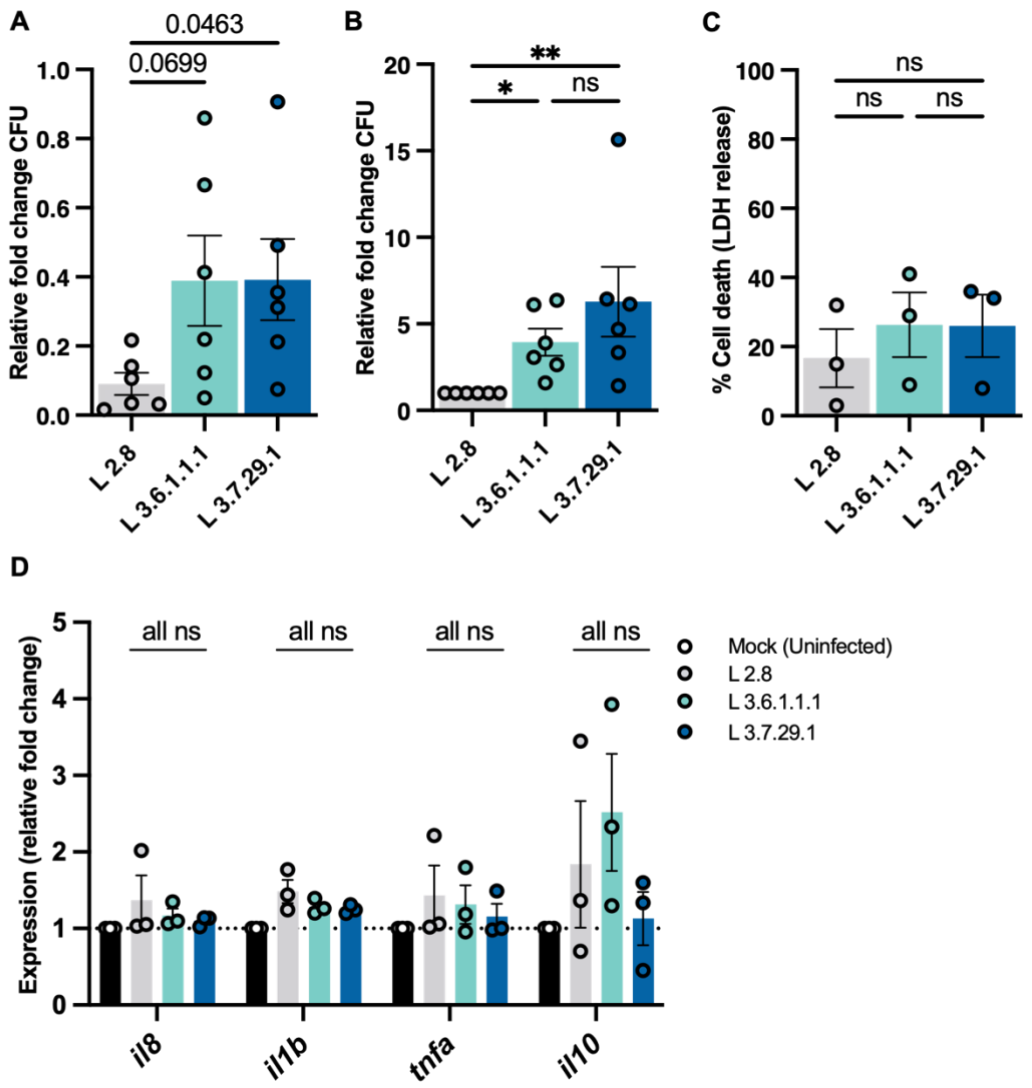


Figure 5.10. Infection of human neutrophils with *S. sonnei*. Human neutrophils were isolated from whole blood and infected in suspension at an MOI of 0.01. Cells were infected for 1 hour before being processed for different assays. **A)** Cells were lysed at 1 hpi and the lysates were plated, colonies counted were normalised to the inoculum to determine relative fold change CFU. **B)** Data presented in panel A is normalised to 53G to account for variation between biological replicates. **C)** Cell death was measured using an LDH release assay at 1 hpi **D)** RNA was extracted from infected/mock neutrophils at 1 hpi and qRT-PCR was performed to analyse gene expression. Relative gene expression was determined using the delta delta Ct method and human *gapdh* was used as a housekeeping gene. Data is representative of 3-6 independent experiments, and all experiments were performed in technical duplicates. Error bars represent the mean ± SEM. Significance was tested using a one- (A,B,C) or two-way ANOVA (D), with Tukey's correction applied. *p<0.0332; **p<0.0021. ns = non-significant, L = lineage.

5.2.4. Lineage 3 *S. sonnei* has an increased stress tolerance *in vitro*

Considering that data from neutrophil infections indicated an enhanced tolerance to neutrophil killing by lineage 3 isolates, I hypothesised that lineage 3 could have an increased tolerance to other stressors. Therefore, I tested the ability to grow under acidic conditions, and in the presence of mammalian complement. First, bacterial growth was measured under normal (pH 7) and acidic conditions (pH 5) to test acid tolerance. At pH 7, there were no significant differences in growth between lineage 2 and 3 isolates over time (Fig 5.11 A), however under acidic conditions, the growth of a lineage 2 isolate was impeded at later time points, compared to lineage 3 (Fig 5.11 B), showing that lineage 3 can grow more efficiently in acidic environments, a property which could contribute to enhanced resistance to phagosomal killing.

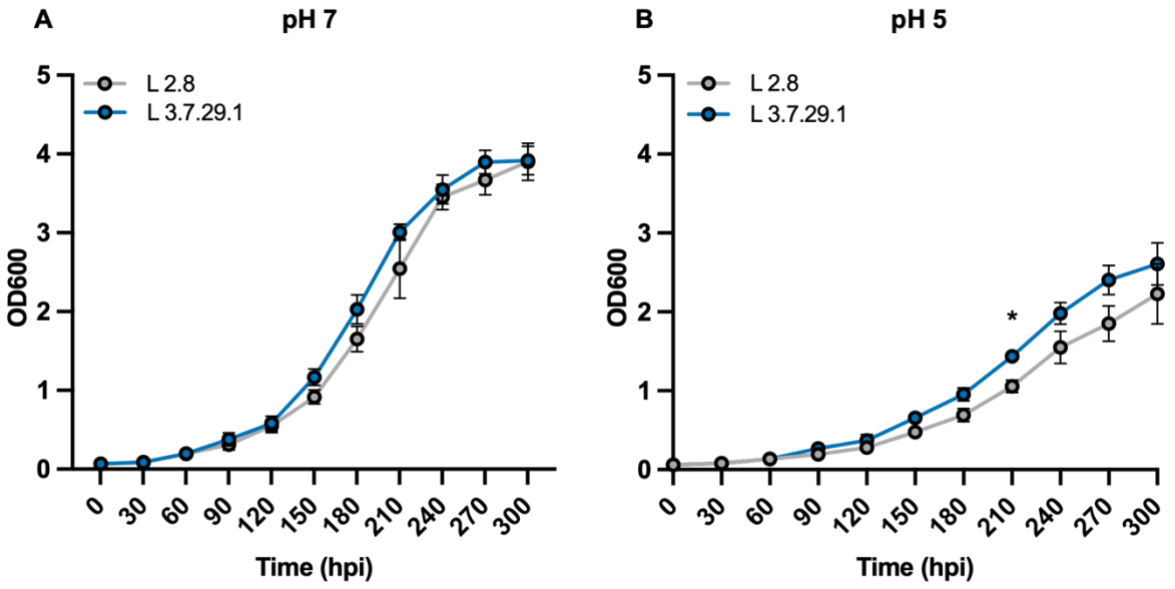


Figure 5.11. Growth of *S. sonnei* lineages under normal and acidic conditions. Bacterial growth curves were performed by inoculating TSB and measuring the optical density (OD) every 30 minutes. **A)** Growth in TSB at pH 7. **B)** Growth in TSB with an adjusted pH of 5, pH was adjusted through the addition of a few drops of concentrated hydrochloric acid. Data is representative of 3 independent experiments, and all experiments were performed in technical duplicates. Error bars represent the mean ± SEM. Significance was tested using a two-way ANOVA, with Tukey's correction applied. *p<0.0332; L = lineage.

Complement-mediated killing is an important component of the innate immune system; its main function is the destruction of foreign cells, but many bacteria have developed strategies to overcome complement-mediated killing (Abreu and Barbosa, 2017). The G4C of *S. sonnei* has previously been shown to aid in resistance to complement mediated killing, but this has only been tested using a lineage 2, lab-adapted isolate (Caboni et al., 2015). Therefore, I next tested whether lineage 3 isolates had differential sensitivity to complement-mediated killing by measuring bacterial growth in the presence of 75% baby rabbit serum. As expected, I detected that the un-capsulated *S. sonnei* mutant was unable to resist the complement system, but both lineage 3 isolates I tested were more tolerant of incubation with complement, with a 2-3.5-fold increase in CFU observed compared to lineage 2.8 (Fig 5.12 A). To ensure observations in bacterial recovery are attributed to complement-mediated killing, baby rabbit serum was heat-inactivated (ensuring denaturation of the complement system) and in this case, there were no differences in bacterial recovery post serum incubation (Fig 5.12 B).

Together, these data show that lineage 3 isolates are more robust under two independent stress conditions, with enhanced growth under acidic conditions and increased resistance to complement-mediated killing.

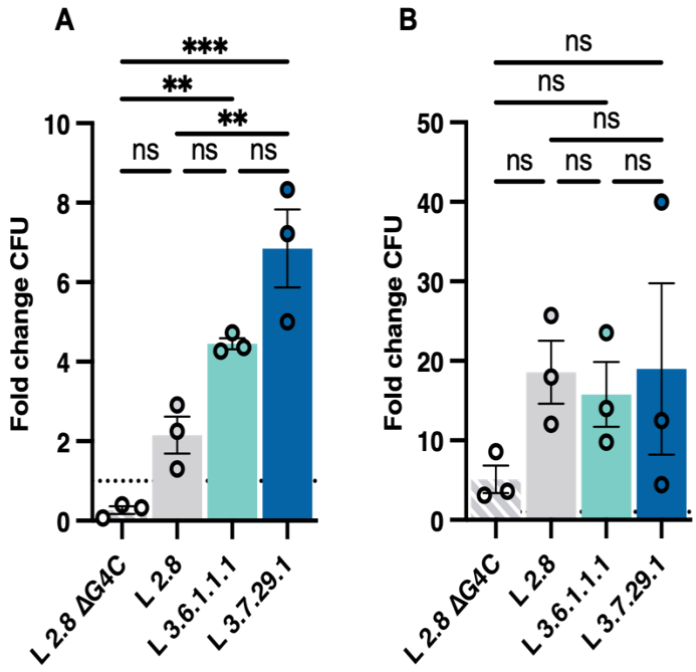


Figure 5.12. Sensitivity of *S. sonnei* lineages to baby rabbit serum. Complement sensitivity was assessed by incubating bacteria in 75% baby rabbit complement and plating the inoculum at 0 and 4 hours. Fold change CFU was determined by normalisation to the 0 hour timepoint. **A)** Fold change CFU in non-treated baby rabbit serum **B)** Fold change CFU in heat-killed baby rabbit serum, serum was incubated at 59 °C for 30 minutes before inoculation to inactivate complement. ΔG4C represents a capsule mutant in L 2.8 and is included as a control. Data is representative of 3 independent experiments, and all experiments were performed in technical duplicates. Error bars represent the mean ± SEM. Significance was tested using a one-way ANOVA, with Tukey's correction applied. **p<0.0021. ns = non-significant, L = lineage.

The O-antigen of *S. sonnei* has been shown to mediate neutrophil tolerance (Torraca et al., 2019), and the G4C has been implicated in resisting complement-mediated killing and dissemination in a rabbit infection model (Caboni et al., 2015). Considering this, I hypothesised that there could be variations in the outer surface layers of lineage 3 *S. sonnei* that may be driving increased virulence and stress response compared to other lineages. G4C synthesis and assembly relies on two gene clusters in *S. sonnei*: one chromosomally encoded cluster which encodes *gfcABCDE*, *etk* and *etp* (Fig 13 A) (Nadler et al., 2012) and is involved in

polysaccharide export; and the pINV encoded *wzy*-dependent O-antigen locus, which is involved in polysaccharide synthesis (Caboni et al., 2015). Genes *etk* and *etp* (which mediate phosphorylation cycling) are essential for G4C formation in enteropathogenic *E. coli* (EPEC) and enterohaemorrhagic *E. coli* (EHEC) (Nadler et al., 2012). To determine whether G4C is differentially regulated in *S. sonnei* lineages, qRT-PCR was performed on bacteria to measure expression levels of core G4C genes *etk* and *etp* (Fig 13 B,C). This revealed a ~1.5-2-fold increase in *etk* expression levels in lineage 3 isolates when compared to lineage 2, and a ~2-3-fold increase in the expression of *etp*, although the difference for *etp* expression was not statistically significant.

Overall, these experiments show the upregulation of at least one capsule export gene in lineage 3 isolates, which could contribute towards the enhanced stress tolerance observed for lineage 3.

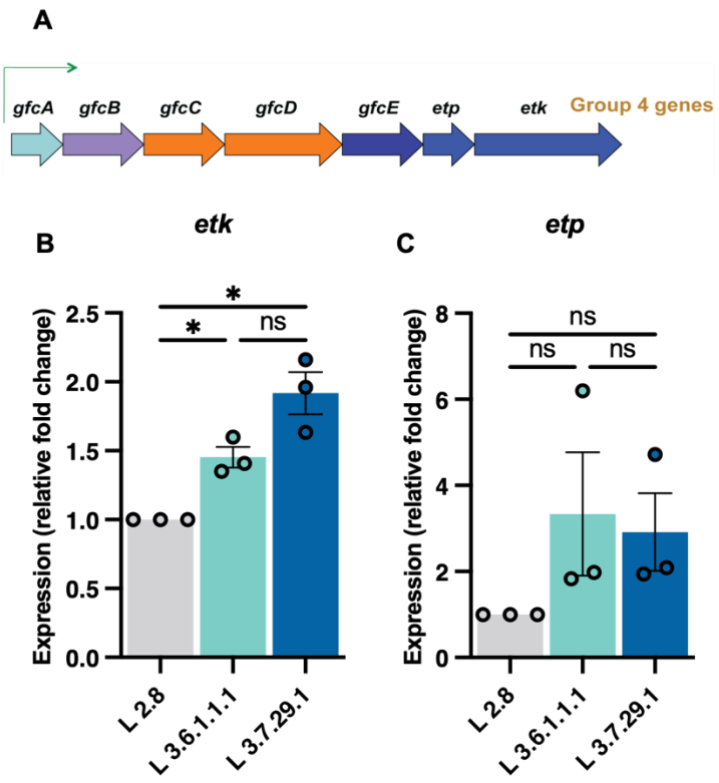


Figure 5.13. Expression of group four capsule (G4C) genes in *S. sonnei* lineages. A) Schematic depicting organisation of the chromosomally encoded G4C formation operon. **B, C)** Relative mRNA expression levels of *etk* **(A)** and *etp* **(B)** in bacteria grown *in vitro*. Bacteria were grown at 37°C in 5 mL TSB, with shaking at 400 rotations per minute until mid-exponential phase (OD 0.5-0.6), where an amount corresponding to $1x10^9$ cells was pelleted prior to RNA isolation. RNA was extracted from pelleted bacteria and qRT-PCR was performed on the resulting cDNA, relative gene expression was determined by normalisation to *rrsA* as a housekeeping gene, and then to L 2.8 using the delta delta Ct method. Data is representative of 3 independent experiments, and all qRT-PCR analysis was performed in technical duplicates. Error bars represent the mean ± SEM. Significance was tested using a one-way ANOVA, with Tukey's correction applied. *p<0.0332. ns= non-significant, L = lineage.

Modulation of the O-antigen chain length has been shown to contribute to virulence in *S. flexneri* (Van den Bosch et al., 1997) and to serum sensitivity in pathogenic *E. coli* (Osawa et al., 2013). In *S. sonnei,* the O-antigen chain is monomodal and consists of ~20 -25 repeating polysaccharide units (Xu et al., 2002), but studies describing this have been restricted to 53G (lineage 2.8). O-antigen chain length is controlled by the product of *wzzB* (Carter et al., 2009). To determine if chain length varies in *S. sonnei* lineages, expression levels of *wzzB* were first

analysed by qRT-PCR. The relative mRNA expression levels of *wzzB* were similar between the lineage 2 and 3 isolates tested (Fig 5.14 B), suggesting no differences in O-antigen chain length. To confirm this, crude LPS was extracted from bacteria, separated by SDS-PAGE and visualised using a previously described modified silver staining protocol (Tsai and Frasch, 1982). This confirmed that all *S. sonnei* lineages tested had a short-chain O-antigen length, with a similar number of repeating units (Fig 5.14 C). Two mutants were included as controls: Δ*waaL*, which is a complete O-antigen mutant, and ΔG4C, a G4C mutant. As expected, no O-antigen staining in Δ*waaL* was detectable, but interestingly there was a much greater O-antigen signal detected in ΔG4C as compared to the wild type lineage 2.8. These results suggest that the regulation of capsule and LPS is a dynamic process, and in the absence of (or reduction of) capsule, *S. sonnei* could compensate for this by upregulating its LPS O-antigen production. Testing of this would require further experimentation and a more quantitative measure of both capsule and LPS abundance.

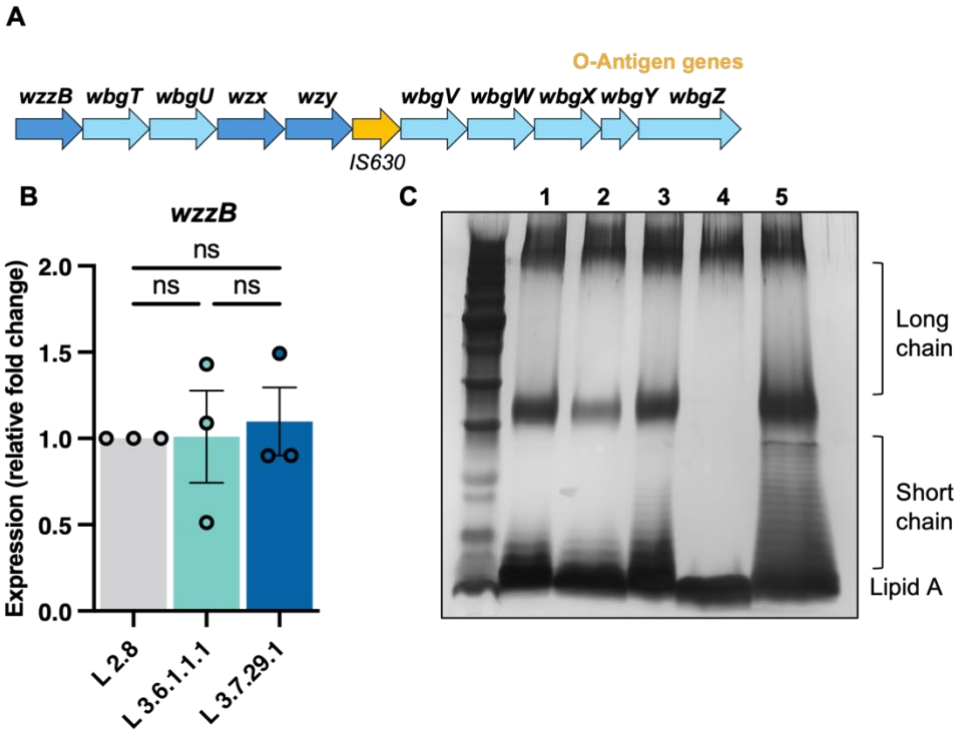


Figure 5.14. Expression of lipopolysaccharide (LPS) in *S. sonnei* lineages. A) Schematic depicting organisation of the pINV encoded O-antigen synthesis gene cluster. **B)** Relative mRNA expression levels of *wzzB* in bacteria grown *in vitro*. Bacteria were grown at 37°C in 5 mL TSB, with shaking at 400 rotations per minute until mid-exponential phase (OD 0.5-0.6), where an amount corresponding to $1x10^9$ cells was pelleted prior to RNA isolation. RNA was extracted from pelleted bacteria and qRT-PCR was performed on the resulting cDNA, relative gene expression was determined by normalisation to *rrsA* as a housekeeping gene, and then to L 2.8 using the delta delta Ct method. Data is representative of 3 independent experiments, and all qRT-PCR analysis was performed in technical duplicates. Error bars represent the mean ± SEM. Significance was tested using a one-way ANOVA, with Tukey's correction applied. ns = non-significant, L = lineage **C)** Visualisation of LPS extracts using SDS-PAGE

5.5. Discussion

5.5.1. Overview

To date, experimental studies on *S. sonnei* have primarily focused on laboratory-adapted isolates, which are no longer representative of the current epidemiological landscape. In this chapter, key *S. sonnei* isolates of epidemiological importance (which were sequenced and compared in Chapter 4) were experimentally characterised to reveal functional differences between successful lineages and those from less prevalent ones.

In this work, I used the zebrafish infection model to test the pathogenicity of different isolates *in vivo*. I found that epidemiologically successful lineage 3 isolates are more virulent, coinciding with more neutrophil recruitment and a greater pro-inflammatory immune response during infection. Infection of primary human neutrophils revealed that lineage 3 *S. sonnei* are more tolerant of neutrophil killing, and further *in vitro* characterisation showed that lineage 3 generally has a higher stress tolerance compared to a lineage 2 isolate. Overall, this chapter delivers a deeper understanding of virulence in *S. sonnei* and highlights both virulence and stress tolerance as key signatures of epidemiologically successful *S. sonnei* isolates.

5.5.2. Lineage 3 *S. sonnei* are more virulent in a zebrafish infection model

Epidemiological success in *S. sonnei* has previously been linked to the carriage of AMR determinants (Holt et al., 2012, Baker et al., 2018b) and interbacterial competition weapons, such as colicins (De Silva et al., 2023a, Leung et al., 2024). Here, zebrafish infection of numerous *S. sonnei* lineages revealed differences in virulence, highlighting for the first time that epidemiologically successful lineage 3 *S. sonnei* is more virulent than lineage 1 and 2. Bacterial replication, pINV stability and T3SS activity were all investigated as potential drivers of virulence but, in both cases, no lineage-dependent differences were identified. Interestingly, lineage 3 isolates were more likely to disseminate from the infection site than lineage 2.8. This phenotype has been demonstrated in zebrafish previously, where *S. sonnei* was found to disseminate at a much greater frequency than *S. flexneri* (Lensen et al., 2023). The implications of this phenotype in zebrafish currently remain unclear but could be indicative of

increased resistance to systemic immune responses. Similarly, in an *S. sonnei* rabbit infection model it was shown that uncapsulated *S. sonnei* could not disseminate as efficiently as its wild type equivalent (Caboni et al., 2015). This could indicate a differential surface composition in lineage 3, which allows it to resist host immune responses and disseminate more efficiently.

Ongoing evolution and genome reduction have been implicated as a predictive marker for increased pathogenicity in a wide range of bacteria (Murray et al., 2021). The genomic data presented in Chapter 4, revealing a reduced genome in lineage 3 isolates, along with the increased virulence reported here, suggests that lineage 3 *S. sonnei* may be evolving toward heightened pathogenicity as it adapts to a more host-restricted lifestyle.

5.5.3. Lineage 3 *S. sonnei* induces a stronger pro-inflammatory immune response

A well-documented step in the evolution *of Shigella* (and other human-adapted pathogens) is the loss of immunogenic components such as flagella and fimbriae (Bravo et al., 2015). The proposed explanation for this is that as new bacterial pathogens adapt to their pathogenic lifestyle, it is more advantageous to avoid immune detection and subsequent destruction. Given the conserved loss of Type 1 and K88 fimbriae-encoding operons in *S. sonnei* lineage 3 (as discussed in Chapter 4), a reduced inflammatory response during infection in these clades might have been speculated. However, results presented here challenge this hypothesis, showing that lineage 3 instead induces a greater upregulation of pro-inflammatory cytokines, and consistent with this, increased neutrophil recruitment to the infection site. These results initially imply that the loss of type 1 and K88 fimbriae do not significantly aid in the immune evasion of *S. sonnei*, but to formally test this, isogenic mutants would need to be created. An alternative explanation for increased inflammation is that lineage 3 *S. sonnei* is not better at evading immune recognition but can alternatively withstand the bactericidal effects of the immune system more efficiently, which ultimately promotes their virulence.

5.5.4. Lineage 3 *S. sonnei* are more tolerant of neutrophil killing *ex vivo*

During its co-evolution with the human host, *Shigella* has evolved many mechanisms to efficiently infect human cells and avoid destruction (Ashida et al., 2015). *S. sonnei* 53G

(lineage 2.8) has previously been shown to be less invasive in HeLa cells than *S. flexneri* (Watson et al., 2019). I hypothesised that *S. sonnei* becomes more proficient at infecting epithelial cells as it continues to evolve, becoming more like *S. flexneri*. However, upon infecting HeLa cells with lineage 2 and 3 *S. sonnei,* no statistical differences in bacterial invasion were observed, and the opposite trend towards decreased invasion in lineage 3 isolates was instead noted. This indicates that increased invasion of epithelial cells is not driving the increased virulence of lineage 3 observed in zebrafish. The G4C has been shown to negatively impact *S. sonnei* invasion (Caboni et al., 2015). A possible explanation for this could be differentially regulated G4C expression or synthesis if lineage 3 is truly less invasive. The capacity of lineage 3 *S. sonnei* to become intracellular deserves further investigation and such experiments may be more suited to a more complex organoid/chip model, which can simulate a more realistic gastric environment and has been shown to increase the invasion of *S. flexneri* (Boquet-Pujadas et al., 2022).

The tolerance of *S. sonnei* to neutrophil killing has been previously documented in the *S. sonnei*-zebrafish model, where the O-antigen was identified as a mediating factor (Torraca et al., 2019). Here, I show that lineage 3 *S. sonnei* can further resist neutrophil killing more than lineage 2.8 but is not more cytotoxic and does not induce greater cytokine expression in neutrophils. This finding therefore implies that lineage 3 *S. sonnei* could have variations in its O-antigen structure, a hypothesis which will be the focus of further studies. Overall, enhanced survival during phagocytosis could benefit the long-term survival and shedding of *S. sonnei*, contributing to the establishment of persistent infections and its onward expansion and epidemiological success, in agreement with previous work (Torraca et al., 2023).

5.5.5. Lineage 3 *S. sonnei* has an increased stress tolerance *in vitro*

Previous *in silico* studies have suggested enhanced adaptation to oxidative stress in a subgroup of epidemiologically successful *S. sonnei* isolates from South Asia (Chung The et al., 2019). Experimental work carried out here is congruent with this hypothesis, showing that epidemiologically successful lineage 3 isolates exhibit enhanced tolerance to growth under

acidic conditions and complement-mediated killing. I hypothesised that differences in stress response could be due to variations in the outer surface components of lineage 3 *S. sonnei*. Gene expression analysis demonstrated a slight upregulation in the expression of capsule synthesis genes but no differences in the expression of LPS chain length modulating genes. However, it is important to recognise the limitations of carrying out such analysis *in vitro*. It is well known that *Shigella* modulates both its virulence and surface polysaccharide properties in response to different environmental cues during infection (Grassart et al., 2019, Van den Bosch et al., 1997). Therefore, it could well be the case that the experiments carried out here have not fully captured this highly dynamic process. Due to the insufficient invasion of *S. sonnei* in HeLa cells, and the complex nature of isolating enough bacterial RNA from an infected whole animal, the study of these properties *in cellulo* and *in vivo* has not been possible to date, but developing these techniques further should be a priority for future work.

5.5.6. Conclusions

In conclusion, these data show for the first time that epidemiologically successful lineage 3 *S. sonnei* isolates are more virulent than their less epidemiologically relevant counterparts. Enhanced virulence is linked to a greater pro-inflammatory response *in vivo*, mediated by increased neutrophil recruitment. Additionally, results indicate an increased stress tolerance in lineage 3 isolates, with increased resistance to neutrophil killing, acidic conditions and complement-mediated killing documented. Further characterisation to pinpoint precise molecular mechanisms underlying these differences will be important for limiting the further spread of successful clones.

Chapter 6. Conclusions and perspectives

6.1. Summary of key findings

The evolutionary history of *Shigella* and EIEC has been characterised by the convergent gain of functions that have promoted virulence, and the process of genome streamlining which has facilitated a unique specialisation to the human host (The et al., 2016, Yang et al., 2010). Characterising the progression of evolutionary adaptation in *Shigella*, and how this impacts subsequent bacterial fitness and pathogenicity, is important to further our understanding of the selection pressures that drive *Shigella* evolution; the niches that various *Shigella* pathovariants may occupy; and to inform targeted preventative and therapeutic measures. Here, a combination of genomic and experimental methodologies was used to characterise ST99 EIEC, a subgroup that is at a relatively early evolutionary stage, and *S. sonnei*, an established *Shigella* subgroup that is responsible for most infections in developed countries and is rapidly replacing other *Shigella* subgroups in developing countries.

My main findings are:

(i) ST99 EIEC diverged from a MRCA ~40 years ago following the acquisition of pINV, illuminating its short evolutionary history (Chapter 3).

(ii) The acquisition of pINV was key in shaping the virulence of ST99 EIEC, but alternative virulence mechanisms were already present in ST99 *E. coli* (Chapter 3).

(iii) The evolution of lineage 3 *S. sonnei* is shaped by the continuing process of genome reduction (Chapter 4).

(iv) The accumulation of ISs in lineage 3 *S. sonnei* is promoting functional changes in the number of pseudogenes and genome rearrangements (Chapter 4).

(v) Epidemiologically successful lineage 3 *S. sonnei* are more virulent in a zebrafish infection model (Chapter 5).

(vi) Lineage 3 *S. sonnei* have a more robust stress response than less epidemiologically relevant lineages (Chapter 5).

6.2. Acquisition of a large virulence plasmid (pINV) promoted temperature-dependent virulence and global dispersal of O96:H19 enteroinvasive *Escherichia coli*

In Chapter 3, ST99 EIEC is used as a model system to study the early stages of *Shigella* emergence. Phylogenomic dating analysis of all publicly available genomes for this ST revealed that the ST99 EIEC pathotype emerged relatively recently, ~40 years ago. Strikingly, zebrafish infection of both EIEC and non-EIEC ST99 isolates uncovered distinct mechanisms: one that is pINV and T3SS dependent, requiring a temperature closer to that of humans; and another that is pINV independent and exists in both EIEC and non-EIEC ST99 isolates. Based on these results, it is proposed that ST99 EIEC was likely circulating undetected for around three decades before its involvement in a European diarrhoeal outbreak in 2012. The presence of pINV-independent virulence mechanisms in both EIEC and non-EIEC ST99 isolates furthermore suggests that alternative virulence mechanisms were already present in ST99 *E. coli*, challenging the hypothesis that *Shigella* and EIEC have evolved from innocuous *E. coli* ancestors. This finding is largely in agreement with hypotheses suggested by Sims and Kim, who propose that *Shigella* and EIEC likely emerged from a uropathogenic *E. coli* (UPEC) strain, given their frequency in the basal phylogroup of the *Shigella* and EIEC phylogeny (Sims and Kim, 2011).

Shigella and EIEC subgroups have arisen independently from *E. coli* following pINV acquisition on at least ten independent occasions (Baker et al., 2024). Considering the most recent subgroup has emerged in the last 50 years, it is clear this process is ongoing, and it is therefore likely that we will see the emergence of novel subgroups again in the future. Future work in this area should consequently aim to first identify the currently unknown source of pINV, and the mechanisms that permit its transfer since most pINV forms are non-conjugative (Yang et al., 2010). Widescale environmental sampling may uncover other pINV-like plasmids that could provide clues as to the source of pINV, for example, the genomic characterisation of *Escherichia marmotae* discovered the presence of a *Shigella*-like T3SS and effectors (Liu et al., 2019) and *Shigella*-like *ipaH* genes have also been detected in *E. coli* from faecal

samples of bovine calves (Dranenko et al., 2022). Additionally, further characterisation of EIEC subgroups and their close relatives will be important to identify traits that may pre-dispose an *E. coli* lineage to acquire pINV and begin the process of human specialisation. The identification of such traits could ultimately aid in forecasting the emergence of novel clones and preventing their widespread dissemination.

6.3. Comparative genomic analysis highlights evidence of ongoing adaptation in lineage 3 *Shigella sonnei*

In Chapter 4, a collection of epidemiologically important *S. sonnei* isolates was established and 15 new pINV-positive genomes were sequenced to completion, which will be deposited to act as a bioresource for the *Shigella* community. A comparative genomic analysis highlighted that epidemiologically successful lineage 3 genomes were smaller, had more ISs and pseudogenes, and had more structural variation compared to lineage 1 and lineage 2 genomes. From these results, it is concluded that the process of reductive evolution is continuing to shape the evolution of *S. sonnei* and it is suggested that the lineage-specific genomic variations highlighted here confer a selective advantage to lineage 3 *S. sonnei* playing a role in its domination of the global epidemiological landscape (Hawkey et al., 2021, Holt et al., 2012). Overall, these results further the current understanding of *S. sonnei* evolution to include variation in IS activity and subsequent structural variation.

A tendency towards genome reduction has been observed in some of humanity's most important bacterial pathogens (Chain et al., 2004, Chavarro-Portillo et al., 2019), but the precise benefits of reductive evolution remain unclear. It would be interesting to compare the variations documented here with other *Shigella* subgroups, and other human adapted pathogens to identify the features that are commonly under selection, which could again aid in forecasting the emergence of pathogens. Alternatively, the identification of differential changes may provide an insight into the unique selection pressures acting on lineage 3 *S. sonnei* and could inform the lifestyle changes that have shaped its evolution. It is possible that the genomic rearrangements described here may confer an advantage for lineage 3

S. sonnei to more readily adapt to changing environments, irrespective of genome reduction. The functional impact of all the genomic variations discovered here was not fully explored in this thesis and future work should aim to determine whether they also drive the changes in pathogenicity described in Chapter 5, perhaps through a wide scale transcriptomic analysis exploring the effect on gene expression.

6.4. Experimental characterisation reveals increased virulence and an increased stress tolerance in lineage 3 *Shigella sonnei*

In Chapter 5, a combination of *in vivo* and *in cellulo* infection models were used to characterise the pathogenicity of epidemiologically relevant *S. sonnei* isolates for the first time. Epidemiological success in *S. sonnei* had previously been linked to the carriage of AMR determinants (Baker et al., 2018b) and bacterial competition weapons (such as colicins and a potential T6SS) (De Silva et al., 2023). Results presented here show that epidemiologically successful lineage 3 isolates are also both more virulent in an *in vivo* model, and more tolerant of neutrophil killing and complement mediated killing. The significant phenotypic variation observed between the lab-adapted lineage 2.8 isolate and more epidemiologically relevant lineage 3 isolates also have important implications for future experimental work, highlighting that just a single lab-adapted isolate should not be used to draw broader conclusions on the pathogenicity of *Shigella.* It would be interesting to link the findings of this thesis to clinical data in future, to determine whether lineage 3 causes more severe disease in humans, but this requires extensive ethical considerations and is beyond the scope of this thesis.

Combined with results from Chapter 4, which are indicative of genome reduction in lineage 3, these experimental results imply that genome reduction may be concurrent with an increase in bacterial pathogenicity, a phenomenon that has been reported for many bacterial pathogens (Murray et al., 2021, Weinert et al., 2015, Merhej et al., 2013). The correlation between reductive genome evolution and bacterial pathogenicity, and any evolutionary advantages that this phenomenon may confer, remain poorly understood (Weinert and Welch, 2017), but

further study and characterisation will be useful for predicting lineages likely to develop enhanced pathogenicity.

Collectively, these findings suggest that compared to less epidemiologically successful *S. sonnei* lineages, lineage 3 has refined the requirements to achieve epidemiological success, through a combination of high carriage rates of stably fixed AMR determinants, enhanced virulence, and being equipped to better resist stressors both in the environment and within the human host.

6.5. Final remarks

This thesis has linked the evolutionary and epidemiological landscape of *Shigella* to both genomic variations and variations in pathogenicity and fitness using a variety of experimental models. This work has delivered novel insights into signatures of epidemiological success in *S. sonnei*, which are characterised by genome reduction, increased virulence and tolerance to stress.

References

Abreu, A. G. & Barbosa, A. S. 2017. How *Escherichia coli* Circumvent Complement-Mediated Killing. *Frontiers in Immunology,* 8, 452.

Al Mamun, A. A., Tominaga, A. & Enomoto, M. 1997. Cloning and characterization of the region III flagellar operons of the four *Shigella* subgroups: genetic defects that cause loss of flagella of *Shigella boydii* and *Shigella sonnei*. *Journal of Bacteriology,* 179, 4493-4500.

Alkindy, B. & Guyeux, C. 2022. Impact of Insertion Sequences and RNAs on Genomic Inversions in *Pseudomonas aeruginosa*. *Journal of King Saud University - Computer and Information Sciences,* 34, 9513-9522.

Amman, F., D'halluin, A., Antoine, R., Huot, L., Bibova, I., Keidel, K., Slupek, S., Bouquet, P., Coutte, L., Caboche, S., Locht, C., Vecerek, B. & Hot, D. 2018. Primary transcriptome analysis reveals importance of IS elements for the shaping of the transcriptional landscape of *Bordetella pertussis*. *RNA Biology,* 15, 967-975.

Anderson, M. C., Vonaesch, P., Saffarian, A., Marteyn, B. S. & Sansonetti, P. J. 2017. *Shigella sonnei* Encodes a Functional T6SS Used for Interbacterial Competition and Niche Occupancy. *Cell Host & Microbe,* 21, 769-776.e3.

Andrews, S. 2010. *FastQC: A Quality Control Tool for High Throughput Sequence Data.* [Online]. Available: http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ [Accessed].

Arena, E. T., Campbell-Valois, F.-X., Tinevez, J.-Y., Nigro, G., Sachse, M., Moya-Nilges, M., Nothelfer, K., Marteyn, B., Shorte, S. L. & Sansonetti, P. J. 2015. Bioimage analysis of *Shigella* infection reveals targeting of colonic crypts. *Proceedings of the National Academy of Sciences of the United States of America* 112, E3282-E3290.

Ashida, H., Mimuro, H. & Sasakawa, C. 2015. *Shigella* manipulates host immune responses by delivering effector proteins with specific roles. *Frontiers in Immunology,* 6, 219.

Baker, K. S., Campos, J., Pichel, M., Della Gaspera, A., Duarte-Martínez, F., Campos-Chacón, E., Bolaños-Acuña, H. M., Guzmán-Verri, C., Mather, A. E., Diaz Velasco, S., Zamudio Rojas, M. L., Forbester, J. L., Connor, T. R., Keddy, K. H., Smith, A. M., López De Delgado, E. A., Angiolillo, G., Cuaical, N., Fernández, J., Aguayo, C., Morales Aguilar, M., Valenzuela, C., Morales Medrano, A. J., Sirok, A., Weiler Gustafson, N., Diaz Guevara, P. L., Montaño, 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. *Clinical Microbiology and Infection,* 23**,** 845-853.

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. *Emerging Infectious Diseases,* 22**,** 1545-53.

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. 2018a. Genomic epidemiology of *Shigella* in the United Kingdom shows transmission of pathogen sublineages and determinants of antimicrobial resistance. *Scientific Reports,* 8**,** 7389.

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., Hawkey, J., Ingle, D., Miles, S. L. & The, H. C. 2024. Chapter 12 - The phylogenomics of *Shigella* spp. *In:* MOKROUSOV, I. & SHITIKOV, E. (eds.) *Phylogenomics.* Academic Press.

Baker, S. & Scott, T. A. 2023. Antimicrobial-resistant *Shigella*: where do we go next? *Nature Reviews Microbiology,* 21**,** 409-410.

Barbagallo, M., Di Martino, M. L., Marcocci, L., Pietrangeli, P., De Carolis, E., Casalino, M., Colonna, B. & Prosseda, G. 2011. A New Piece of the *Shigella* Pathogenicity Puzzle: Spermidine Accumulationby Silencing of the *speG* Gene. *PLOS ONE,* 6**,** e27226.

Barnhart, M. M. & Chapman, M. R. 2006. Curli biogenesis and function. *Annual Review of Microbiology.,* 60**,** 131-147.

Bernardini, M. L., Mounier J., D'hauteville, H., Coquis-Rondon, M., Coquis-Rondon & Sansonetti, P. J. 1989. Identification of *icsA*, a plasmid locus of *Shigella flexneri* that governs bacterial intra- and intercellular spread through interaction with F-actin. *Proceedings of the National Academy of Sciences of the United States of America*, *86*(10), 3867–3871.

Better, M. 1999. 4 - *araB* expression system in *Escherichia coli*. *In:* FERNANDEZ, J. M. & HOEFFLER, J. P. (eds.) *Gene Expression Systems.* San Diego: Academic Press.

Bliven, K. A. & Maurelli, A. T. 2012. Antivirulence genes: Insights into pathogen evolution through gene loss. *Infection and Immunity,* 80**,** 4061-4070.

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-2120.

Boquet-Pujadas, A., Feaugas, T., Petracchini, A., Grassart, A., Mary, H., Manich, M., Gobaa, S., Olivo-Marin, J.-C., Sauvonnet, N. & Labruyère, E. 2022. 4D live imaging and computational modeling of a functional gut-on-a-chip evaluate how peristalsis facilitates enteric pathogen invasion. *Science Advances,* 8**,** eabo5767.

Boucontet, L., Passoni, G., Thiry, V., Maggi, L., Herbomel, P., Levraud, J. P. & Colucci-Guyon, E. 2018. A Model of Superinfection of Virus-Infected Zebrafish Larvae: Increased Susceptibility to Bacteria Associated With Neutrophil Death. *Frontiers in Immunology,* 9**,** 1084.

Bouras, G., Grigson, S. R., Papudeshi, B., Mallawaarachchi, V. & Roach, M. 2024a. Dnaapler: A tool to reorient circular microbial genomes. *The Journal of Open Source Software,* 9**,** 5968.

Bouras, G., Houtak, G., Wick, R. R., Mallawaarachchi, V., Roach, M. J., Papudeshi, B., Judd, L. M., Sheppard, A. E., Edwards, R. A. & Vreugde, S. 2024b. Hybracter: enabling scalable, automated, complete and accurate bacterial genome assemblies. *Microbial Genomics,* 10.

Bouras, G., Judd, L. M., Edwards, R. A., Vreugde, S., Stinear, T. P. & Wick, R. R. 2024c. How low can you go? Short-read polishing of Oxford Nanopore bacterial genome assemblies. *Microbial Genomics,* 10.

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.

Brokatzky, D., Gomes, M. C., Robertin, S., Albino, C., Miles, S. L. & Mostowy, S. 2024. Septins promote macrophage pyroptosis by regulating gasdermin D cleavage and ninjurin-1-mediated plasma membrane rupture. *Cell Chemical Biology,* 31**,** 1518-1528.e6.

Brotcke Zumsteg, A., Goosmann, C., Brinkmann, V., Morona, R. & Zychlinsky, A. 2014. IcsA is a *Shigella flexneri* adhesin regulated by the type III secretion system and required for pathogenesis. *Cell Host & Microbe,* 15**,** 435-45.

Caboni, M. 2013. *Identification and characterization of a capsule in Shigella and its role in the pathogenesis.* Università degli Studi di Napoli Federico II Open Archive.

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 Pathogens,* 11(3): e1004749

Calcuttawala, F., Hariharan, C., Pazhani, G. P., Saha, D. R. & Ramamurthy, T. 2017. Characterization of E-type colicinogenic plasmids from *Shigella sonnei*. *FEMS Microbiology Letters,* 364.

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.

Carter, J. A., Jiménez, J. C., Zaldívar, M., Álvarez, S. A., Marolda, C. L., Valvano, M. A. & Contreras, I. 2009. The cellular level of O-antigen polymerase Wzy determines chain length regulation by WzzB and WzzpHS-2 in *Shigella flexneri* 2a. *Microbiology,* 155**,** 3260-3269.

Casalino, M., Latella Mc Fau - Prosseda, G., Prosseda G Fau - Colonna, B. & Colonna, B. 2003. CadC is the preferential target of a convergent evolution driving enteroinvasive *Escherichia coli* toward a lysine decarboxylase-defective phenotype. *Infection and immunity*, *71*(10), 5472–5479.

Chain, P. S., Carniel, E., Larimer, F. W., Lamerdin, J., Stoutland, P. O., Regala, W. M., Georgescu, A. M., Vergez, L. M., Land, M. L., Motin, V. L., Brubaker, R. R., Fowler, J., Hinnebusch, J., Marceau, M., Medigue, C., Simonet, M., Chenal-Francisque, V., Souza, B., Dacheux, D., Elliott, J. M., Derbise, A., Hauser, L. J. & Garcia, E. 2004. Insights into the evolution of *Yersinia pestis* through whole-genome comparison with *Yersinia pseudotuberculosis*. *Proceedings of the National Academy of Sciences of the United States of America,* 101**,** 13826-31.

Chandler, M. & Mahillon, J. 2007. Insertion sequences revisited. *Mobile DNA II***,** 303-366.

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., Corkin, H., Abrahams, A., Leblond, H., Lo, J., Holgate, A., Saunders, J., Plahe, G., Vusirikala, A., Green, F., King, M., Tewolde, R. & Jajja, A. 2022. Outbreak

of sexually transmitted, extensively drug-resistant *Shigella sonnei* in the UK, 2021–22: a descriptive epidemiological study. *The Lancet Infectious Diseases,* 22**,** 1503-1510.

Chavarro-Portillo, B., Soto, C. Y. & Guerrero, M. I. 2019. *Mycobacterium leprae*'s evolution and environmental adaptation. *Acta Tropica,* 197**,** 105041.

Chung The, H. & Baker, S. 2018. Out of Asia: the independent rise and global spread of fluoroquinolone-resistant *Shigella*. *Microbial Genomics,* 4(4):e000171.

Chung The, H., Bodhidatta, L., Pham, D. T., Mason, C. J., Ha Thanh, T., Voong Vinh, P., Turner, P., Hem, S., Dance, D. a. B., Newton, P. N., Phetsouvanh, R., Davong, V., Thwaites, G. E., Thomson, N. R., Baker, S. & Rabaa, M. A. 2021. Evolutionary histories and antimicrobial resistance in *Shigella flexneri* and *Shigella sonnei* in Southeast Asia. *Communications Biology,* 4**,** 353.

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., Bodhidatta, L., Mason, C. J., Nguyen, T. N. T., Ha Thanh, T., Voong, V. P., Duong, V. T., Nguyen, P. H. L., Turner, P., Wick, R., Ceyssens, P.-J., Thwaites, G., Holt, K. E., Thomson, N. R., Rabaa, M. A. & Baker, S. 2019. Dissecting the molecular evolution of fluoroquinolone-resistant *Shigella sonnei*. *Nature Communications,* 10**,** 4828.

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 Med,* 13**,** e1002055.

Connor, T. R., Barker, C. R., Baker, K. S., Weill, F.-X., Talukder, K. A., Smith, A. M., Baker, S., Gouali, M., Pham Thanh, D. & Jahan Azmi, I. 2015. Species-wide whole genome sequencing reveals historical global spread and recent local persistence in *Shigella flexneri*. *elife,* 4**,** e07335.

Darling, A. E., Mau, B. & Perna, N. T. 2010. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. *PLOS One,* 5**,** e11147.

Darling, A. E., Miklós, I. & Ragan, M. A. 2008. Dynamics of Genome Rearrangement in Bacterial Populations. *PLOS Genetics,* 4**,** e1000128.

Davis, M. R., Jr. & Goldberg, J. B. 2012. Purification and visualization of lipopolysaccharide from Gram-negative bacteria by hot aqueous-phenol extraction. *Journal of Visualized Experiments*.

Day, W. A., Jr., Fernández, R. E. & Maurelli, A. T. 2001. Pathoadaptive mutations that enhance virulence: genetic organization of the cadA regions of *Shigella* spp. *Infection and Immunity,* 69**,** 7471-80.

De Silva, P. M., Bennett, R. J., Kuhn, L., Ngondo, P., Debande, L., Njamkepo, E., Ho, B., Weill, F. X., Marteyn, B. S., Jenkins, C. & Baker, K. S. 2023. *Escherichia coli* killing by epidemiologically successful sublineages of *Shigella sonnei* is mediated by colicins. *EBioMedicine,* 97**,** 104822.

Di Martino, M. L., Falconi, M., Micheli, G., Colonna, B. & Prosseda, G. 2016. The Multifaceted Activity of the VirF Regulatory Protein in the *Shigella* Lifestyle. *Frontiers in Molecular Biosciences,* 3**,** 61.

Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. 2018. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Research*, *46*(22), e134.

Diop, A., Raoult, D. & Fournier, P.-E. 2018. Rickettsial genomics and the paradigm of genome reduction associated with increased virulence. *Microbes and Infection,* 20**,** 401-409.

Dranenko, N. O., Tutukina, M. N., Gelfand, M. S., Kondrashov, F. A. & Bochkareva, O. O. 2022. Chromosome-encoded IpaH ubiquitin ligases indicate non-human enteroinvasive *Escherichia*. *Scientific Reports,* 12**,** 6868.

Duarte-Velázquez, I., De La Mora, J., Ramírez-Prado, J. H., Aguillón-Bárcenas, A., Tornero-Gutiérrez, F., Cordero-Loreto, E., Anaya-Velázquez, F., Páramo-Pérez, I., Rangel-Serrano, Á., Muñoz-Carranza, S. R., Romero-González, O. E., Cardoso-Reyes, L. R., Rodríguez-Ojeda, R. A., Mora-Montes, H. M., Vargas-Maya, N. I., Padilla-Vaca, F. & Franco, B. 2022. *Escherichia coli* transcription factors of unknown function: sequence features and possible evolutionary relationships. *PeerJ,* 10**,** e13772.

Duchêne, S., Holt, K. E., Weill, F. X., Le Hello, S., Hawkey, J., Edwards, D. J., Fourment, M. & Holmes, E. C. 2016. Genome-scale rates of evolutionary change in bacteria. *Microbial Genomics,* 2**,** e000094.

Dupont, H. L., Formal, S. B., Hornick, R. B., Snyder, M. J., Libonati, J. P., Sheahan, D. G., Labrec, E. H. & Kalas, J. P. 1971. Pathogenesis of *Escherichia coli* diarrhea. *New England Journal of Medicine,* 285**,** 1-9.

Dupont, H. L., Levine, M. M., Hornick, R. B. & Formal, S. B. 1989. Inoculum Size in Shigellosis and Implications for Expected Mode of Transmission. *The Journal of Infectious Diseases,* 159**,** 1126-1128.

Ellett, F., Pase, L., Hayman, J. W., Andrianopoulos, A. & Lieschke, G. J. 2011. *mpeg1* promoter transgenes direct macrophage-lineage expression in zebrafish. *Blood,* 117**,** e49-56.

Erridge, C., Bennett-Guerrero, E. & Poxton, I. R. 2002. Structure and function of lipopolysaccharides. *Microbes and infection,* 4**,** 837-851.

Escher, M., Scavia, G., Morabito, S., Tozzoli, R., Maugliani, A., Cantoni, S., Fracchia, S., Bettati, A., Casa, R., Gesu, G. P., Torresani, E. & Caprioli, A. 2014. A severe foodborne outbreak of diarrhoea linked to a canteen in Italy caused by enteroinvasive *Escherichia coli*, an uncommon agent. *Epidemiology and Infection,* 142**,** 2559-2566.

Ewing, W. & Gravatti, J. 1947. *Shigella* types encountered in the mediterranean area. *Journal of Bacteriology,* 53**,** 191-195.

Falconi, M., Colonna, B., Prosseda, G., Micheli, G. & Gualerzi, C. O. 1998. Thermoregulation of *Shigella* and *Escherichia coli* EIEC pathogenicity. A temperature-dependent structural transition of DNA modulates accessibility of *virF* promoter to transcriptional repressor H-NS. *The The EMBO Journalournal,* 17**,** 7033-7043.

Feldgarden, M., Brover, V., Gonzalez-Escalona, N., Frye, J. G., Haendiges, J., Haft, D. H., Hoffmann, M., Pettengill, J. B., Prasad, A. B., Tillman, G. E., Tyson, G. H. & Klimke, W. 2021. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. *Scientific Reports,* 11**,** 12728.

Fullá, N., Prado, V., Durán, C., Lagos, R. & Levine, M. M. 2005. Surveillance for antimicrobial resistance profiles among *Shigella* species isolated from a semirural community in the northern administrative area of Santiago, Chile. *The American journal of tropical medicine and hygiene,* 72**,** 851-854.

Gilchrist, C. L. M. & Chooi, Y.-H. 2021. clinker & clustermap.js: automatic generation of gene cluster comparison figures. *Bioinformatics,* 37**,** 2473-2475.

Goldberg, M. B. & Theriot, J. A. 1995. *Shigella flexneri* surface protein IcsA is sufficient to direct actin-based motility. *Proceedings of the National Academy of Sciences of the United States of America,* 92**,** 6572-6.

Goldsmith, J. R. & Jobin, C. 2012. Think small: zebrafish as a model system of human pathology. *Journal of Biomedicine and Biotechnology,* 2012**,** 817341.

Gomes, M. C., Brokatzky, D., Bielecka, M. K., Wardle, F. C. & Mostowy, S. 2023. *Shigella* induces epigenetic reprogramming of zebrafish neutrophils. *Science Advances,* 9**,** eadf9706.

Gomes, T. A., Elias, W. P., Scaletsky, I. C., Guth, B. E., Rodrigues, J. F., Piazza, R. M., Ferreira, L. & Martinez, M. B. 2016. Diarrheagenic *Escherichia coli*. *Brazilian journal of microbiology,* 47**,** 3-30.

Grassart, A., Malardé, V., Gobaa, S., Sartori-Rupp, A., Kerns, J., Karalis, K., Marteyn, B., Sansonetti, P. & Sauvonnet, N. 2019. Bioengineered Human Organ-on-Chip Reveals Intestinal Microenvironment and Mechanical Forces Impacting *Shigella* Infection. *Cell Host & Microbe,* 26**,** 435-444.e4.

Guan, S., Bastin, D. A. & Verma, N. K. 1999. Functional analysis of the O antigen glucosylation gene cluster of *Shigella flexneri* bacteriophage SfX. *Microbiology,* 145**,** 1263-1273.

Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. 2013. QUAST: quality assessment tool for genome assemblies. *Bioinformatics,* 29**,** 1072-5.

Guyet, A., Dade-Robertson, M., Wipat, A., Casement, J., Smith, W., Mitrani, H. & Zhang, M. 2018. Mild hydrostatic pressure triggers oxidative responses in *Escherichia coli.* PLOS ONE, 13, e0200660.

Hachani, A., Biskri, L., Rossi, G., Marty, A., Ménard, R., Sansonetti, P., Parsot, C., Van Nhieu, G. T., Bernardini, M. L. & Allaoui, A. 2008. IpgB1 and IpgB2, two homologous effectors secreted via the Mxi-Spa type III secretion apparatus, cooperate to mediate polarized cell invasion and inflammatory potential of *Shigella flexneri*. *Microbes and Infection,* 10**,** 260-268.

Hadfield, J., Croucher, N. J., Goater, R. J., Abudahab, K., Aanensen, D. M. & Harris, S. R. 2017. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics,* 34**,** 292-293.

Hasegawa, T., Hall, C. J., Crosier, P. S., Abe, G., Kawakami, K., Kudo, A. & Kawakami, A. 2017. Transient inflammatory response mediated by interleukin-1β is required for proper regeneration in zebrafish fin fold. *eLife,* 6.

Hawkey, J., Monk, J., Billman-Jacobe, H., Palsson, B. & Holt, K. 2020. Impact of insertion sequences on convergent evolution of *Shigella* species, PLOS Genetics, 16(7):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.

Hazen, T., H., Leonard, S., R., Lampel, K., A., Lacher, D., W., Maurelli, A., T. & Rasko, D., A. 2016. Investigating the Relatedness of Enteroinvasive *Escherichia coli* to Other *E. coli* and *Shigella* Isolates by Using Comparative Genomics. *Infection and Immunity,* 84**,** 2362-2371.

Herbst, S., Shah, A., Mazon Moya, M., Marzola, V., Jensen, B., Reed, A., Birrell, M. A., Saijo, S., Mostowy, S., Shaunak, S., & Armstrong-James, D. (2015). Phagocytosis-dependent activation of a TLR9-BTK-calcineurin-NFAT pathway co-ordinates innate immunity to *Aspergillus fumigatus*. *EMBO molecular medicine*, *7*, 240–258.

Hershberg, R., Tang, H. & Petrov, D. A. 2007. Reduced selection leads to accelerated gene loss in *Shigella*. *Genome Biology,* 8**,** R164.

High, N., Mounier, J., Prévost, M. C. & Sansonetti, P. J. 1992. IpaB of *Shigella flexneri* causes entry into epithelial cells and escape from the phagocytic vacuole. *The EMBO Journal,* 11**,** 1991-9.

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. 2012. *Shigella sonnei* genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nature Genetics,* 44**,** 1056-9.

Holt, K. E., Thieu Nga, T. V., Thanh, D. P., Vinh, H., Kim, D. W., Vu Tra, M. P., 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 of the United States of America,* 110**,** 17522-17527.

Howe, K., Clark, M. D., Torroja, C. F., Torrance, J., Berthelot, C., Muffato, M., Collins, J. E., Humphray, S., Mclaren, K., Matthews, L., Mclaren, S., Sealy, I., Caccamo, M., Churcher, C., Scott, C., Barrett, J. C., Koch, R., Rauch, G.-J., White, S., Chow, W., Kilian, B., Quintais, L. T., Guerra-Assunção, J. A., Zhou, Y., Gu, Y., Yen, J., Vogel, J.-H., Eyre, T., Redmond, S., Banerjee, R., Chi, J., Fu, B., Langley, E., Maguire, S. F., Laird, G. K., Lloyd, D., Kenyon, E., Donaldson, S., Sehra, H., Almeida-King, J., Loveland, J., Trevanion, S., Jones, M., Quail, M., Willey, D., Hunt, A., Burton, J., Sims, S., Mclay, K., Plumb, B., Davis, J., Clee, C., Oliver, K., Clark, R., Riddle, C., Elliott, D., Threadgold, G., Harden, G., Ware, D., Begum, S., Mortimore, B., Kerry, G., Heath, P., Phillimore, B., Tracey, A., Corby, N., Dunn, M., Johnson, C., Wood, J., Clark, S., Pelan,

S., Griffiths, G., Smith, M., Glithero, R., Howden, P., Barker, N., Lloyd, C., Stevens, C., Harley, J., Holt, K., Panagiotidis, G., Lovell, J., Beasley, H., Henderson, C., Gordon, D., Auger, K., Wright, D., Collins, J., Raisen, C., Dyer, L., Leung, K., Robertson, L., Ambridge, K., Leongamornlert, D., Mcguire, S., Gilderthorp, R., Griffiths, C., Manthravadi, D., Nichol, S., Barker, G., Whitehead, S., Kay, M., Brown, J., Murnane, C., Gray, E., Humphries, M., Sycamore, N., Barker, D., Saunders, D., Wallis, J., Babbage, A., Hammond, S., Mashreghi-Mohammadi, M., Barr, L., Martin, S., Wray, P., Ellington, A., Matthews, N., Ellwood, M., Woodmansey, R., Clark, G., Cooper, J. D., Tromans, A., Grafham, D., Skuce, C., Pandian, R., Andrews, R., Harrison, E., Kimberley, A., Garnett, J., Fosker, N., Hall, R., Garner, P., Kelly, D., Bird, C., Palmer, S., Gehring, I., Berger, A., Dooley, C. M., Ersan-Ürün, Z., Eser, C., Geiger, H., Geisler, M., Karotki, L., Kirn, A., Konantz, J., Konantz, M., Oberländer, M., Rudolph-Geiger, S., Teucke, M., Lanz, C., Raddatz, G., Osoegawa, K., Zhu, B., Rapp, A., Widaa, S., Langford, C., Yang, F., Schuster, S. C., Carter, N. P., Harrow, J., Ning, Z., Herrero, J., Searle, S. M. J., Enright, A., Geisler, R., Plasterk, R. H. A., Lee, C., Westerfield, M., De Jong, P. J., Zon, L. I., Postlethwait, J. H., Nüsslein-Volhard, C., Hubbard, T. J. P., Crollius, H. R., Rogers, J. & Stemple, D. L. 2013. The zebrafish reference genome sequence and its relationship to the human genome. *Nature,* 496**,** 498-503.

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. 2019. Antibiotic resistance prediction for *Mycobacterium tuberculosis* from genome sequence data with Mykrobe. *Wellcome Open Research,* 4**,** 191.

Ibañez, E., Campos, E., Baldoma, L., Aguilar, J. & Badia, J. 2000. Regulation of expression of the *yiaKLMNOPQRS* operon for carbohydrate utilization in *Escherichia coli*: involvement of the main transcriptional factors. *Journal of Bacteriology,* 182**,** 4617-24.

Jin, Q., Yuan, Z., Xu, J., Wang, Y., Shen, Y., Lu, W., Wang, J., Liu, H., Yang, J., Yang, F., Zhang, X., Zhang, J., Yang, G., Wu, H., Qu, D., Dong, J., Sun, L., Xue, Y., Zhao, A., Gao, Y., Zhu, J., Kan, B., Ding, K., Chen, S., Cheng, H., Yao, Z., He, B., Chen, R., Ma, D., Qiang, B., Wen, Y., Hou, Y. & Yu, J. 2002. Genome sequence of *Shigella flexneri* 2a: insights into pathogenicity through comparison with genomes of *Escherichia coli* K12 and O157. *Nucleic Acids Research,* 30**,** 4432-4441.

Jolley, K. A., Bray, J. E. & Maiden, M. C. J. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Research,* 3**,** 124.

Karaolis, D., Lan, R. & Reeves, P. R. 1994. Sequence variation in *Shigella sonnei*, a pathogenic clone of *Escherichia coli*, over four continents and 41 years. *Journal of clinical microbiology,* 32**,** 796-802.

Katoh, K. & Standley, D. M. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution,* 30**,** 772-80.

Khalil, I., Troeger, C. E., Blacker, B. F. & Reiner, R. C. 2019. Capturing the true burden of *Shigella* and ETEC: The way forward. *Vaccine,* 37**,** 4784-4786.

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., Kotloff, K. L., Levine, M. M., Luby, S. P.,

Maclennan, C. A., Pan, W. K., Pavlinac, P. B., Platts-Mills, J. A., Qadri, F., Riddle, M. S., Ryan, E. T., Shoultz, D. A., Steele, A. D., Walson, J. L., Sanders, J. W., Mokdad, A. H., Murray, C. J. L., Hay, S. I. & Reiner, R. C., Jr. 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. 1986. Two regulatory fim genes, *fimB* and *fimE*, control the phase variation of type 1 fimbriae in *Escherichia coli*. *The The EMBO Journal,* 5**,** 1389-1393.

Kolmogorov, M., Yuan, J., Lin, Y. & Pevzner, P. A. 2019. Assembly of long, error-prone reads using repeat graphs. *Nature Biotechnology,* 37**,** 540-546.

Koppolu, V., Osaka, I., Skredenske, J. M., Kettle, B., Hefty, P. S., Li, J. & Egan, S. M. 2013. Small-molecule inhibitor of the *Shigella flexneri* master virulence regulator VirF. *Infection and Immunity,* 81**,** 4220-31.

Koren, S. & Phillippy, A. M. 2015. One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly. *Current Opinion in Microbiology,* 23**,** 110-120.

Kotloff, K. L., Blackwelder, W. C., Nasrin, D., Nataro, J. P., Farag, T. H., Van Eijk, A., Adegbola, R. A., Alonso, P. L., Breiman, R. F., Faruque, A. S., Saha, D., Sow, S. O., Sur, D., Zaidi, A. K., Biswas, K., Panchalingam, S., Clemens, J. D., Cohen, D., Glass, R. I., Mintz, E. D., Sommerfelt, H. & Levine, M. M. 2012. The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study. *Clinical Infectious Diseases,* 55 Suppl 4**,** S232-45.

Kotloff, K. L., Riddle, M. S., Platts-Mills, J. A., Pavlinac, P. & Zaidi, A. K. M. 2018. Shigellosis. *The Lancet,* 391**,** 801-812.

Kotloff, K. L., Winickoff Jp Fau - Ivanoff, B., Ivanoff B Fau - Clemens, J. D., Clemens Jd Fau - Swerdlow, D. L., Swerdlow Dl Fau - Sansonetti, P. J., Sansonetti Pj Fau - Adak, G. K., Adak Gk Fau - Levine, M. M. & Levine, M. M. 1999. Global burden of *Shigella* infections: implications for vaccine development and implementation of control strategies. *Bulletin of the World Health Organization*, *77*(8), 651–666.

Krokowski, S., Lobato-Márquez, D., Chastanet, A., Pereira, P. M., Angelis, D., Galea, D., Larrouy-Maumus, G., Henriques, R., Spiliotis, E. T., Carballido-López, R. & Mostowy, S. 2018. Septins Recognize and Entrap Dividing Bacterial Cells for Delivery to Lysosomes. *Cell Host & Microbe,* 24**,** 866-874.e4.

Kugelberg, E., Gollan, B., Farrance, C., Bratcher, H., Lucidarme, J., Ibarz-Pavón, A. B., Maiden, M. C., Borrow, R. & Tang, C. M. 2010. The influence of IS*1301* in the capsule biosynthesis locus on meningococcal carriage and disease. *PLOS One,* 5**,** e9413.

Lagerqvist, N., Löf, E., Enkirch, T., Nilsson, P., Roth, A. & Jernberg, C. 2020. Outbreak of gastroenteritis highlighting the diagnostic and epidemiological challenges of enteroinvasive *Escherichia coli*, County of Halland, Sweden, November 2017. *Euro surveillance : bulletin Europeen sur les maladies transmissibles = European communicable disease bulletin,* 25**,** 1900466.

Lan, R., Alles, M. C., Donohoe, K., Martinez, M. B. & Reeves, P. R. 2004. Molecular evolutionary relationships of enteroinvasive *Escherichia coli* and *Shigella* spp. *Infection and immunity,* 72**,** 5080-5088.

Lan, R., Lumb, B., Ryan, D. & Reeves, P. R. 2001. Molecular Evolution of Large Virulence Plasmid in *Shigella* Clones and Enteroinvasive *Escherichia coli*. *Infection and Immunity,* 69**,** 6303-6309.

Landfald, B. & Strøm, A. R. 1986. Choline-glycine betaine pathway confers a high level of osmotic tolerance in *Escherichia coli*. *Journal of Bacteriology,* 165**,** 849-855.

Langridge, G. C., Fookes, M., Connor, T. R., Feltwell, T., Feasey, N., Parsons, B. N., Seth-Smith, H. M., Barquist, L., Stedman, A. & Humphrey, T. 2015. Patterns of genome evolution that have accompanied host adaptation in *Salmonella*. *Proceedings of the National Academy of Sciences of the United States of America,* 112**,** 863-868.

Le Guyader, D., Redd, M. J., Colucci-Guyon, E., Murayama, E., Kissa, K., Briolat, V., Mordelet, E., Zapata, A., Shinomiya, H. & Herbomel, P. 2008. Origins and unconventional behavior of neutrophils in developing zebrafish. *Blood, The Journal of the American Society of Hematology,* 111**,** 132-141.

Lensen, A., Gomes, M. C., López-Jiménez, A. T. & Mostowy, S. 2023. An automated microscopy workflow to study *Shigella*–neutrophil interactions and antibiotic efficacy in vivo. *Disease Models & Mechanisms,* 16.

Letunic, I. & Bork, P. 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Research,* 49**,** W293-W296.

Leung, P. B., Matanza, X. M., Roche, B., Ha, K. P., Cheung, H. C., Appleyard, S., Collins, T., Flanagan, O., Marteyn, B. S. & Clements, A. 2024. *Shigella sonnei* utilises colicins during inter-bacterial competition. *Microbiology ,* 170.

Li, Z., Liu, W., Fu, J., Cheng, S., Xu, Y., Wang, Z., Liu, X., Shi, X., Liu, Y., Qi, X., Liu, X., Ding, J. & Shao, F. 2021. *Shigella* evades pyroptosis by arginine ADP-riboxanation of caspase-11. *Nature,* 599**,** 290-295.

Liu, B., Zheng, D., Zhou, S., Chen, L. & Yang, J. 2022. VFDB 2022: a general classification scheme for bacterial virulence factors. *Nucleic Acids Research,* 50**,** D912-d917.

Liu, S., Feng, J., Pu, J., Xu, X., Lu, S., Yang, J., Wang, Y., Jin, D., Du, X., Meng, X., Luo, X., Sun, H., Xiong, Y., Ye, C., Lan, R. & Xu, J. 2019. Genomic and molecular characterisation of *Escherichia marmotae* from wild rodents in Qinghai-Tibet plateau as a potential pathogen. *Scientific Reports,* 9**,** 10619.

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. *Microbial Genomics,* 7.

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. *Infection and Immunity,* 69**,** 6012-21.

Luria, S. E. & Burrous, J. W. 1957. Hybridization between *Escherichia coli* and *Shigella*. *Journal of Bacteriology,* 74**,** 461-76.

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*. *Proceedings of the Royal Society B: Biological Sciences,* 289**,** 20220581.

Marteyn, B., West, N. P., Browning, D. F., Cole, J. A., Shaw, J. G., Palm, F., Mounier, J., Prévost, M.-C., Sansonetti, P. & Tang, C. M. 2010. Modulation of *Shigella* virulence in response to available oxygen in vivo. *Nature,* 465**,** 355-358.

Martinić, M., Hoare, A., Contreras, I. & Álvarez, S. A. 2011. Contribution of the lipopolysaccharide to resistance of *Shigella flexneri* 2a to extreme acidity. *PLOS One,* 6**,** e25557.

Martyn, J. E., Pilla, G., Hollingshead, S., Winther, K. S., Lea, S., Mcvicker, G. & Tang, C. M. 2022. Maintenance of the *Shigella sonnei* Virulence Plasmid Is Dependent on Its Repertoire and Amino Acid Sequence of Toxin-Antitoxin Systems. *Journal of Bacteriology,* 204**,** e0051921.

Mason, L. C. E., Greig, D. R., Cowley, L. A., Partridge, S. R., Martinez, E., Blackwell, G. A., Chong, C. E., De Silva, P. M., Bengtsson, R. J., Draper, J. L., Ginn, A. N., Sandaradura, I., Sim, E. M., Iredell, J. R., Sintchenko, V., Ingle, D. J., Howden, B. P., Lefèvre, S., Njamkepo, E., Weill, F.-X., Ceyssens, P.-J., Jenkins, C. & Baker, K. S. 2023. The evolution and international spread of extensively drug resistant *Shigella sonnei. Nature Communications,* 14**,** 1983.

Maurelli, A. T., Fernández, R. E., Bloch, C. A., Rode, C. K. & Fasano, A. 1998. "Black holes" and bacterial pathogenicity: a large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive *Escherichia coli*. *Proceedings of the National Academy of Sciences of the United States of America,* 95**,** 3943-3948.

Maurelli, A. T. & Sansonetti, P. J. 1988. Identification of a chromosomal gene controlling temperature-regulated expression of *Shigella* virulence. *Proceedings of the National Academy of Sciences of the United States of America,* 85**,** 2820-2824.

Mazon-Moya, M. J., Willis, A. R., Torraca, V., Boucontet, L., Shenoy, A. R., Colucci-Guyon, E. & Mostowy, S. 2017. Septins restrict inflammation and protect zebrafish larvae from *Shigella* infection. *PLOS Pathogens,* 13**,** e1006467.

Merhej, V., Georgiades, K. & Raoult, D. 2013. Postgenomic analysis of bacterial pathogens repertoire reveals genome reduction rather than virulence factors. *Briefings in Functional Genomics,* 12**,** 291-304.

Michelacci, V., Prosseda, G., Maugliani, A., Tozzoli, R., Sanchez, S., Herrera-León, S., Dallman, T., Jenkins, C., Caprioli, A. & Morabito, S. 2016. Characterization of an emergent clone of enteroinvasive *Escherichia coli* circulating in Europe.

Miles, S. L., Torraca, V., Dyson, Z. A., López-Jiménez, A. T., Foster-Nyarko, E., Lobato-Márquez, D., Jenkins, C., Holt, K. E. & Mostowy, S. 2023. Acquisition of a large virulence plasmid (pINV) promoted temperature-dependent virulence and global dispersal of O96:H19 enteroinvasive *Escherichia coli*. *mBio,* 14**,** e00882-23.

Mitchell, P. S., Roncaioli, J. L., Turcotte, E. A., Goers, L., Chavez, R. A., Lee, A. Y., Lesser, C. F., Rauch, I. & Vance, R. E. 2020. NAIP–NLRC4-deficient mice are susceptible to shigellosis. *eLife,* 9**,** e59022.

Mol, O., Oudhuis, W. C., Oud, R. P., Sijbrandi, R., Luirink, J., Harms, N. & Oudega, B. 2001. Biosynthesis of K88 fimbriae in *Escherichia coli*: interaction of tip-subunit FaeC with

the periplasmic chaperone FaeE and the outer membrane usher FaeD. *Journal of Molecular Microbiology and Biotechnology,* 3**,** 135-42.

Moran, N. A. & Plague, G. R. 2004. Genomic changes following host restriction in bacteria. *Current Opinion in Genetics & Development,* 14**,** 627-633.

Moreno, A. C. R., Gonçalves Ferreira, L. & Baquerizo Martinez, M. 2009. Enteroinvasive *Escherichia coli* vs. *Shigella flexneri* : how different patterns of gene expression affect virulence. *FEMS Microbiology Letters,* 301**,** 156-163.

Morita, T., Majdalani, N., Miura, M. C., Inose, R., Oshima, T., Tomita, M., Kanai, A. & Gottesman, S. 2022. Identification of Attenuators of Transcriptional Termination: Implications for RNA Regulation in *Escherichia coli*. *mBio,* 13**,** e0237122.

Moss, J. E., Cardozo, T. J., Zychlinsky, A. & Groisman, E. A. 1999. The *selC*-associated SHI-2 pathogenicity island of *Shigella flexneri*. *Molecular Microbiology,* 33**,** 74-83.

Mostowy, S., Bonazzi, M., Hamon, M. A., Tham, T. N., Mallet, A., Lelek, M., Gouin, E., Demangel, C., Brosch, R., Zimmer, C., Sartori, A., Kinoshita, M., Lecuit, M. & Cossart, P. 2010. Entrapment of intracytosolic bacteria by septin cage-like structures. *Cell Host & Microbe,* 8**,** 433-44.

Mostowy, S., Sancho-Shimizu, V., Hamon, M. A., Simeone, R., Brosch, R., Johansen, T. & Cossart, P. 2011. p62 and NDP52 Proteins Target Intracytosolic *Shigella* and *Listeria* to Different Autophagy Pathways. *Journal of Biological Chemistry,* 286**,** 26987-26995.

Moussouni, M., Berry, L., Sipka, T., Nguyen-Chi, M. & Blanc-Potard, A.-B. 2021. *Pseudomonas aeruginosa* OprF plays a role in resistance to macrophage clearance during acute infection. *Scientific Reports,* 11**,** 359.

Murray, G. G. R., Charlesworth, J., Miller, E. L., Casey, M. J., Lloyd, C. T., Gottschalk, M., Tucker, A. W. D., Welch, J. J. & Weinert, L. A. 2021. Genome Reduction Is Associated with Bacterial Pathogenicity across Different Scales of Temporal and Ecological Divergence. *Molecular Biology and Evolution,* 38**,** 1570-1579.

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.

Nadler, C., Koby, S., Peleg, A., Johnson, A. C., Suddala, K. C., Sathiyamoorthy, K., Smith, B. E., Saper, M. A. & Rosenshine, I. 2012. Cycling of Etk and Etp phosphorylation states is involved in formation of group 4 capsule by *Escherichia coli*. *PLOS One,* 7**,** e37984.

Nakayama, S.-I. & Watanabe, H. 1995. Involvement of *cpxA*, a sensor of a two-component regulatory system, in the pH-dependent regulation of expression of *Shigella sonnei* virF gene. *Journal of Bacteriology,* 177**,** 5062-5069.

Newitt, S., Macgregor, V., Robbins, V., Bayliss, L., Chattaway, M. A., Dallman, T., Ready, D., Aird, H., Puleston, R. & Hawker, J. 2016. Two Linked Enteroinvasive *Escherichia coli* Outbreaks, Nottingham, UK, June 2014. *Emerging Infectious Diseases,* 22**,** 1178-1184.

Niesel, D. W., Chambers, C. E. & Stockman, S. L. 1985. Quantitation of HeLa cell monolayer invasion by *Shigella* and *Salmonella* species. *Journal of Clinical Microbiology,* 22**,** 897-902.

Njamkepo, E., Fawal, N., Tran-Dien, A., Hawkey, J., Strockbine, N., Jenkins, C., Talukder, K. A., Bercion, R., Kuleshov, K. & Kolínská, R. 2016. Global phylogeography and evolutionary history of *Shigella dysenteriae* type 1. *Nature Microbiology,* 1**,** 1-10.

Ochman, H. & Moran, N. A. 2001. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science,* 292**,** 1096-1099.

Oehlers, S. H. B., Flores, M. V., Hall, C. J., O'toole, R., Swift, S., Crosier, K. E. & Crosier, P. S. 2010. Expression of zebrafish *cxcl8* (interleukin-8) and its receptors during development and in response to immune stimulation. *Developmental & Comparative Immunology,* 34**,** 352-359.

Ogasawara, H., Yamamoto, K. & Ishihama, A. 2011. Role of the biofilm master regulator CsgD in cross-regulation between biofilm formation and flagellar synthesis. *Journal of Bacteriology,* 193**,** 2587-97.

Osawa, K., Shigemura, K., Iguchi, A., Shirai, H., Imayama, T., Seto, K., Raharjo, D., Fujisawa, M., Osawa, R. & Shirakawa, T. 2013. Modulation of O-antigen chain length by the *wzz* gene in *Escherichia coli* O157 influences its sensitivities to serum complement. *Microbiology and Immunology,* 57**,** 616-23.

Palić, D., Andreasen, C. B., Ostojić, J., Tell, R. M. & Roth, J. A. 2007. Zebrafish (*Danio rerio*) whole kidney assays to measure neutrophil extracellular trap release and degranulation of primary granules. *Journal of Immunological Methods,* 319**,** 87-97.

Parajuli, P., Adamski, M. & Verma, N. K. 2017. Bacteriophages are the major drivers of *Shigella flexneri* serotype 1c genome plasticity: a complete genome analysis. *BMC Genomics,* 18**,** 722.

Park, J., Zhang, Y., Buboltz, A. M., Zhang, X., Schuster, S. C., Ahuja, U., Liu, M., Miller, J. F., Sebaihia, M., Bentley, S. D., Parkhill, J. & Harvill, E. T. 2012. Comparative genomics of the classical *Bordetella* subspecies: the evolution and exchange of virulence-associated diversity amongst closely related pathogens. *BMC Genomics,* 13**,** 545.

Parkhill, J., Sebaihia, M., Preston, A., Murphy, L. D., Thomson, N., Harris, D. E., Holden, M. T. G., Churcher, C. M., Bentley, S. D., Mungall, K. L., Cerdeño-Tárraga, A. M., Temple, L., James, K., Harris, B., Quail, M. A., Achtman, M., Atkin, R., Baker, S., Basham, D., Bason, N., Cherevach, I., Chillingworth, T., Collins, M., Cronin, A., Davis, P., Doggett, J., Feltwell, T., Goble, A., Hamlin, N., Hauser, H., Holroyd, S., Jagels, K., Leather, S., Moule, S., Norberczak, H., O'neil, S., Ormond, D., Price, C., Rabbinowitsch, E., Rutter, S., Sanders, M., Saunders, D., Seeger, K., Sharp, S., Simmonds, M., Skelton, J., Squares, R., Squares, S., Stevens, K., Unwin, L., Whitehead, S., Barrell, B. G. & Maskell, D. J. 2003. Comparative analysis of the genome sequences of *Bordetella pertussis*, *Bordetella parapertussis* and *Bordetella bronchiseptica*. *Nature Genetics,* 35**,** 32-40.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research,* 25**,** 1043-55.

Parsot, C. 2009. *Shigella* type III secretion effectors: how, where, when, for what purposes? *Current Opinion in Microbiology,* 12**,** 110-116.

Payne, D., O'reilly, M. & Williamson, D. 1993. The K88 fimbrial adhesin of enterotoxigenic *Escherichia coli* binds to beta 1-linked galactosyl residues in glycosphingolipids. *Infection and Immunity,* 61**,** 3673-7.

Peirano, V., Bianco, M. N., Navarro, A., Schelotto, F. & Varela, G. 2018. Diarrheagenic *Escherichia coli* Associated with Acute Gastroenteritis in Children from Soriano, Uruguay. *Canadian Journal of Infectious Diseases and Medical Microbiology,* 2018**,** 8387218.

Peng, J., Yang, J. & Jin, Q. 2009. The molecular evolutionary history of *Shigella* spp. and enteroinvasive *Escherichia coli*. *Infection, Genetics and Evolution,* 9**,** 147-152.

Phillippy, A. M., Schatz, M. C. & Pop, M. 2008. Genome assembly forensics: finding the elusive mis-assembly. *Genome Biology,* 9**,** 1-13.

Pinske, C. 2018. The Ferredoxin-Like Proteins HydN and YsaA Enhance Redox Dye-Linked Activity of the Formate Dehydrogenase H Component of the Formate Hydrogenlyase Complex. *Frontiers in Microbiology,* 9**,** 1238.

Pomposiello, P. J., Bennik, M. H. & Demple, B. 2001. Genome-wide transcriptional profiling of the *Escherichia coli* responses to superoxide stress and sodium salicylate. *Journal of Bacteriology,* 183**,** 3890-902.

Price, M. N., Dehal, P. S. & Arkin, A. P. 2010. FastTree 2--approximately maximum-likelihood trees for large alignments. *PLOS One,* 5**,** e9490.

Prosseda, G., Di Martino, M. L., Campilongo, R., Fioravanti, R., Micheli, G., Casalino, M. & Colonna, B. 2012. Shedding of genes that interfere with the pathogenic lifestyle: the *Shigella* model. *Research in Microbiology,* 163**,** 399-406.

Prosseda, G., Falconi, M., Giangrossi, M., Gualerzi, C. O., Micheli, G. & Colonna, B. 2004. The *virF* promoter in *Shigella*: more than just a curved DNA stretch. *Molecular Microbiology,* 51**,** 523-537.

Prunier, A. L., Schuch R Fau - Fernández, R. E., Fernández Re Fau - Maurelli, A. T. & Maurelli, A. T. 2007. Genetic structure of the *nadA* and *nadB* antivirulence loci in *Shigella* spp. *Journal of Bacteriology*, *189*(17), 6482–6486.

Pulford, C. V., Perez-Sepulveda, B. M., Canals, R., Bevington, J. A., Bengtsson, R. J., Wenner, N., Rodwell, E. V., Kumwenda, B., Zhu, X., Bennett, R. J., Stenhouse, G. E., Malaka De Silva, P., Webster, H. J., Bengoechea, J. A., Dumigan, A., Tran-Dien, A., Prakash, R., Banda, H. C., Alufandika, L., Mautanga, M. P., Bowers-Barnard, A., Beliavskaia, A. Y., Predeus, A. V., Rowe, W. P. M., Darby, A. C., Hall, N., Weill, F.-X., Gordon, M. A., Feasey, N. A., Baker, K. S. & Hinton, J. C. D. 2021. Stepwise evolution of *Salmonella* Typhimurium ST313 causing bloodstream infection in Africa. *Nature Microbiology,* 6**,** 327-338.

Pupo, G. M., Karaolis, D., Lan, R. & Reeves, P. R. 1997. Evolutionary relationships among pathogenic and nonpathogenic *Escherichia coli* strains inferred from multilocus enzyme electrophoresis and mdh sequence studies. *Infection and immunity,* 65**,** 2685-2692.

Pupo, G. M., Lan, R. & Reeves, P. R. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proceedings*

of the National Academy of Sciences of the United States of America, 97**,** 10567-10572.

Purdy, G. E. & Payne, S. M. 2001. The SHI-3 iron transport island of *Shigella boydii* 0-1392 carries the genes for aerobactin synthesis and transport. *Journal of Bacteriology,* 183**,** 4176-82.

Qu, F., Bao C Fau - Chen, S., Chen S Fau - Cui, E., Cui E Fau - Guo, T., Guo T Fau - Wang, H., Wang H Fau - Zhang, J., Zhang J Fau - Wang, H., Wang H Fau - Tang, Y.-W., Tang Yw Fau - Mao, Y. & Mao, Y. 2012. Genotypes and antimicrobial profiles of *Shigella sonnei* isolates from diarrheal patients circulating in Beijing between 2002 and 2007. *Diagnostic Microbiology and Infectious Disease*, *74*(2), 166–170.

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.

Randow, F., Macmicking, J. D. & James, L. C. 2013. Cellular Self-Defense: How Cell-Autonomous Immunity Protects Against Pathogens. *Science,* 340**,** 701-706.

Ravan, H. & Amandadi, M. 2015. Analysis of *yeh* Fimbrial Gene Cluster in *Escherichia coli* O157:H7 in Order to Find a Genetic Marker for this Serotype. *Current Microbiology,* 71**,** 274-282.

Renshaw, S. A., Loynes, C. A., Trushell, D. M. I., Elworthy, S., Ingham, P. W. & Whyte, M. K. B. 2006. A transgenic zebrafish model of neutrophilic inflammation. *Blood,* 108**,** 3976-3978.

Renshaw, S. A. & Trede, N. S. 2012. A model 450 million years in the making: zebrafish and vertebrate immunity. *Disease Models & Mechanisms,* 5**,** 38-47.

Robertson, J. & Nash, J. H. E. 2018. MOB-suite: software tools for clustering, reconstruction and typing of plasmids from draft assemblies. *Microbial Genomics,* 4.

Rojas-Lopez, M., Gil-Marqués, M. L., Kharbanda, V., Zajac, A. S., Miller, K. A., Wood, T. E., Hachey, A. C., Egger, K. T. & Goldberg, M. B. 2023. NLRP11 is a pattern recognition receptor for bacterial lipopolysaccharide in the cytosol of human macrophages. *Science Immunology,* 8**,** eabo4767.

Sack, D. A., Hoque at Fau - Huq, A., Huq a Fau - Etheridge, M. & Etheridge, M. 1994. Is protection against shigellosis induced by natural infection with *Plesiomonas shigelloides*? *Lancet, 343*(8910), 1413–1415.

Saeed, A., Abd, H., Edvinsson, B. & Sandström, G. 2008. *Acanthamoeba castellanii* an environmental host for *Shigella dysenteriae* and *Shigella sonnei*. *Archives of Microbiology,* 191**,** 83-8.

Sakellaris, H., Hannink Nerissa, K., Rajakumar, K., Bulach, D., Hunt, M., Sasakawa, C. & Adler, B. 2000. Curli Loci of *Shigella* spp. *Infection and Immunity,* 68**,** 3780-3783.

Sansonetti, P. J., Kopecko, D. J. & Formal, S. B. 1981. *Shigella sonnei* plasmids: evidence that a large plasmid is necessary for virulence. *Infection and immunity,* 34**,** 75-83.

Sansonetti, P. J., Kopecko, D. J. & Formal, S. B. 1982. Involvement of a plasmid in the invasive ability of *Shigella flexneri*. *Infection and Immunity,* 35**,** 852-60.

Schneider, D., Duperchy, E., Coursange, E., Lenski, R. E. & Blot, M. 2000. Long-term experimental evolution in *Escherichia coli*. IX. Characterization of insertion sequence-mediated mutations and rearrangements. *Genetics,* 156**,** 477-88.

Schneider, E., Freundlieb, S., Tapio, S. & Boos, W. 1992. Molecular characterization of the MalT-dependent periplasmic alpha-amylase of *Escherichia coli* encoded by *malS*. *Journal of Biological Chemistry,* 267**,** 5148-5154.

Schnupf, P. & Sansonetti, P., J. 2019. *Shigella* Pathogenesis: New Insights through Advanced Methodologies. *Microbiology Spectrum,* 7**,** 10.1128/microbiolspec.bai-0023-2019.

Schroeder, G. N. & Hilbi, H. 2008. Molecular pathogenesis of *Shigella* spp.: controlling host cell signaling, invasion, and death by type III secretion. *Clinical Microbiology Reviews*, 21(1):134-56.

Schwengers, O., Jelonek, L., Dieckmann, M. A., Beyvers, S., Blom, J. & Goesmann, A. 2021. Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. *Microbial Genomics,* 7.

Seemann, T. 2016. *Abricate* [Online]. Github. Available: https://github.com/tseemann/abricate [Accessed 2022].

Seferbekova, Z., Zabelkin, A., Yakovleva, Y., Afasizhev, R., Dranenko, N. O., Alexeev, N., Gelfand, M. S. & Bochkareva, O. O. 2021. High Rates of Genome Rearrangements and Pathogenicity of *Shigella* spp. *Frontiers in Microbiology,* 12**,** 628622.

Shen, W., Le, S., Li, Y. & Hu, F. 2016. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLOS ONE,* 11**,** e0163962.

Shepherd, J. G., Wang, L. & Reeves, P. R. 2000. Comparison of O-Antigen Gene Clusters of *Escherichia coli (Shigella) Sonnei* and *Plesiomonas shigelloides* O17: Sonnei Gained Its Current Plasmid-Borne O-Antigen Genes from *P. shigelloides* in a Recent Event. *Infection and Immunity,* 68**,** 6056-6061.

Siguier, P., Gourbeyre, E. & Chandler, M. 2014. Bacterial insertion sequences: their genomic impact and diversity. *FEMS Microbiol Rev,* 38**,** 865-91.

Sims, G. E. & Kim, S.-H. 2011. Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (FFPs). *Proceedings of the National Academy of Sciences of the United States of America,* 108**,** 8329-8334.

Skovajsová, E., Colonna, B., Prosseda, G., Sellin, M. E. & Di martino, M. L. 2022. The VirF21:VirF30 protein ratio is affected by temperature and impacts *Shigella flexneri* host cell invasion. *FEMS Microbiology Letters,* 369.

Small, P., Isberg, R. & Falkow, S. 1987. Comparison of the ability of enteroinvasive *Escherichia coli, Salmonella typhimurium, Yersinia pseudotuberculosis,* and *Yersinia enterocolitica* to enter and replicate within HEp-2 cells. *Infection and immunity,* 55**,** 1674-1679.

Song, S. & Park, C. 1997. Organization and regulation of the D-xylose operons in *Escherichia coli* K-12: XylR acts as a transcriptional activator. *Journal of Bacteriology,* 179**,** 7025-32.

Sousa, C., De Lorenzo, V. & Cebolla, A. 1997. Modulation of gene expression through chromosomal positioning in *Escherichia coli*. *Microbiology,* 143**,** 2071-2078.

Sousa, M., Mendes En Fau - Collares, G. B., Collares Gb Fau - Péret-Filho, L. A., Péret-Filho La Fau - Penna, F. J., Penna Fj Fau - Magalhães, P. P. & Magalhães, P. P. 2013. *Shigella* in Brazilian children with acute diarrhoea: prevalence, antimicrobial resistance and virulence genes. *Memorias do Instituto Oswaldo Cruz*, *108*(1), 30–35.

Stamatakis, A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics,* 30**,** 1312-3.

Stinear, T. P., Seemann, T., Harrison, P. F., Jenkin, G. A., Davies, J. K., Johnson, P. D., Abdellah, Z., Arrowsmith, C., Chillingworth, T. & Churcher, C. 2008. Insights from the complete genome sequence of *Mycobacterium marinum* on the evolution of *Mycobacterium tuberculosis*. *Genome Research,* 18**,** 729-741.

Stockhammer, O. W., Zakrzewska, A., Hegedûs, Z., Spaink, H. P. & Meijer, A. H. 2009. Transcriptome profiling and functional analyses of the zebrafish embryonic innate immune response to *Salmonella* infection. *Journal of Immunology,* 182**,** 5641-53.

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 Research,* 44**,** 6614-24.

Thanh Duy, P., Thi Nguyen, T. N., Vu Thuy, D., Chung The, H., Alcock, F., Boinett, C., Dan Thanh, H. N., Thanh Tuyen, H., Thwaites, G. E., Rabaa, M. A. & Baker, S. 2020. Commensal *Escherichia coli* are a reservoir for the transfer of XDR plasmids into epidemic fluoroquinolone-resistant *Shigella sonnei*. *Nature Microbiology,* 5**,** 256-264.

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.

Thomas, P. D., Ebert, D., Muruganujan, A., Mushayahama, T., Albou, L.-P. & Mi, H. 2022. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Science,* 31**,** 8-22.

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 Neglected Tropical Diseases,* 9**,** e0003708.

Tobe, T., Yoshikawa, M., Mizuno, T. & Sasakawa, C. 1993. Transcriptional control of the invasion regulatory gene *virB* of *Shigella flexneri*: activation by *virF* and repression by H-NS. *Journal of Bacteriology,* 175**,** 6142-9.

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.

Torraca, V., Brokatzky, D., Miles, S. L., Chong, C. E., De Silva, P. M., Baker, S., Jenkins, C., Holt, K. E., Baker, K. S. & Mostowy, S. 2023. *Shigella* Serotypes Associated With Carriage in Humans Establish Persistent Infection in Zebrafish. *The Journal of Infectious Diseases,* 228**,** 1108-1118.

Torraca, V., Kaforou, M., Watson, J., Duggan, G. M., Guerrero-Gutierrez, H., Krokowski, S., Hollinshead, M., Clarke, T. B., Mostowy, R. J., Tomlinson, G. S., Sancho-Shimizu, V., Clements, A. & Mostowy, S. 2019. *Shigella sonnei* infection of zebrafish reveals that O-antigen mediates neutrophil tolerance and dysentery incidence. *PLOS Pathogens,* 15**,** e1008006.

Torraca, V., Otto, N. A., Tavakoli-Tameh, A. & Meijer, A. H. 2017. The inflammatory chemokine Cxcl18b exerts neutrophil-specific chemotaxis via the promiscuous chemokine receptor Cxcr2 in zebrafish. *Developmental & Comparative Immunology,* 67**,** 57-65.

Tsai, C.-M. & Frasch, C. E. 1982. A sensitive silver stain for detecting lipopolysaccharides in polyacrylamide gels. *Analytical Biochemistry,* 119**,** 115-119.

Turner, S. A., Luck, S. N., Sakellaris, H., Rajakumar, K. & Adler, B. 2003. Molecular epidemiology of the SRL pathogenicity island. *Antimicrobial Agents and Chemotherapy,* 47**,** 727-34.

Van Den Beld, M. J. C. & Reubsaet, F. a. G. 2012. Differentiation between *Shigella*, enteroinvasive *Escherichia coli* (EIEC) and noninvasive *Escherichia coli*. *European Journal of Clinical Microbiology & Infectious Diseases,* 31**,** 899-904.

Van Den Bosch, L., Manning, P. A. & Morona, R. 1997. Regulation of O-antigen chain length is required for *Shigella flexneri* virulence. *Molecular Microbiology,* 23**,** 765-75.

Vargas, M., Gascon, J., De Anta, M. T. J. & 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.

Vinh, H., Nhu, N. T. K., Nga, T. V. T., Duy, P. T., Campbell, J. I., Hoang, N. V. M., Boni, M. F., My, P. V. T., Parry, C., Nga, T. T. T., Van Minh, P., Thuy, C. T., Diep, T. S., Phuong, L. T., Chinh, M. T., Loan, H. T., Tham, N. T. H., Lanh, M. N., Mong, B. L., Anh, V. T. C., Bay, P. V. B., Chau, N. V. V., Farrar, J. & Baker, S. 2009. A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation. *BMC Infectious Diseases,* 9**,** 204.

Virgo, M., Mostowy, S. & Ho, B. T. 2024. Use of zebrafish to identify host responses specific to type VI secretion system mediated interbacterial antagonism. *PLOS Pathogens,* 20**,** e1012384.

Vokes, S. A., Reeves, S. A., Torres, A. G. & Payne, S. M. 1999. The aerobactin iron transport system genes in *Shigella flexneri* are present within a pathogenicity island. *Molecular Microbiology,* 33**,** 63-73.

Waldminghaus, T., Heidrich, N., Brantl, S. & Narberhaus, F. 2007. FourU: a novel type of RNA thermometer in *Salmonella*. *Molecular Microbiology,* 65**,** 413-24.

Wang, M. D., Liu, L., Wang, B. M. & Berg, C. M. 1987. Cloning and characterization of the *Escherichia coli* K-12 alanine-valine transaminase (*avtA*) gene. *Journal of Bacteriology,* 169**,** 4228-34.

Waters, E. V., Tucker, L. A., Ahmed, J. K., Wain, J. & Langridge, G. C. 2022. Impact of *Salmonella* genome rearrangement on gene expression. *Evolution Letters,* 6**,** 426-437.

Watson, J., Jenkins, C. & Clements, A. 2018. *Shigella sonnei* Does Not Use Amoebae as Protective Hosts. *Applied and Environmental Microbiology,* 84**,** e02679-17.

Watson, J. L., Sanchez-Garrido, J., Goddard, P. J., Torraca, V., Mostowy, S., Shenoy, A. R. & Clements, A. 2019. *Shigella sonnei* O-Antigen Inhibits Internalization, Vacuole Escape, and Inflammasome Activation. *mBio,* 10.

Wei, J., Goldberg, M. B., Burland, V., Venkatesan, M. M., Deng, W., Fournier, G., Mayhew, G. F., Plunkett, G., 3rd, Rose, D. J., Darling, A., Mau, B., Perna, N. T., Payne, S. M., Runyen-Janecky, L. J., Zhou, S., Schwartz, D. C. & Blattner, F. R. 2003. Complete genome sequence and comparative genomics of *Shigella flexneri* serotype 2a strain 2457T. *Infection and Immunity,* 71**,** 2775-86.

Weigand, M. R., Peng, Y., Loparev, V., Batra, D., Bowden, K. E., Burroughs, M., Cassiday, P. K., Davis, J. K., Johnson, T., Juieng, P., Knipe, K., Mathis, M. H., Pruitt, A. M., Rowe, L., Sheth, M., Tondella, M. L. & Williams, M. M. 2017. The History of *Bordetella pertussis* Genome Evolution Includes Structural Rearrangement. *Journal of Bacteriology,* 199.

Weinert, L. A., Chaudhuri, R. R., Wang, J., Peters, S. E., Corander, J., Jombart, T., Baig, A., Howell, K. J., Vehkala, M., Välimäki, N., Harris, D., Chieu, T. T. B., Van Vinh Chau, N., Campbell, J., Schultsz, C., Parkhill, J., Bentley, S. D., Langford, P. R., Rycroft, A. N., Wren, B. W., Farrar, J., Baker, S., Hoa, N. T., Holden, M. T. G., Tucker, A. W., Maskell, D. J., Bossé, J. T., Li, Y., Maglennon, G. A., Matthews, D., Cuccui, J., Terra, V. & Consortium, B. R. T. 2015. Genomic signatures of human and animal disease in the zoonotic pathogen *Streptococcus suis*. *Nature Communications,* 6**,** 6740.

Weinert, L. A. & Welch, J. J. 2017. Why Might Bacterial Pathogens Have Small Genomes? *Trends in Ecology & Evolution,* 32**,** 936-947.

Wells, T. J., Sherlock, O., Rivas, L., Mahajan, A., Beatson, S. A., Torpdahl, M., Webb, R. I., Allsopp, L. P., Gobius, K. S., Gally, D. L. & Schembri, M. A. 2008. EhaA is a novel autotransporter protein of enterohemorrhagic *Escherichia coli* O157:H7 that contributes to adhesion and biofilm formation. *Environmental Microbiology,* 10**,** 589-604.

Wen, J., Won, D. & Fozo, E. M. 2014. The ZorO-OrzO type I toxin–antitoxin locus: repression by the OrzO antitoxin. *Nucleic Acids Research,* 42**,** 1930-1946.

Wick, R. R. 2017. Filtlong. [Online]. Github. Available: https://github.com/rrwick/Filtlong [Accessed 2024].

Wick, R. R., Judd, L. M., Cerdeira, L. T., Hawkey, J., Méric, G., Vezina, B., Wyres, K. L. & Holt, K. E. 2021. Trycycler: consensus long-read assemblies for bacterial genomes. *Genome Biology,* 22**,** 266.

Wick, R. R., Judd, L. M. & Holt, K. E. 2023. Assembling the perfect bacterial genome using Oxford Nanopore and Illumina sequencing. *PLOS Computational Biology,* 19**,** e1010905.

Williams, K. P. 2002. Integration sites for genetic elements in prokaryotic tRNA and tmRNA genes: sublocation preference of integrase subfamilies. *Nucleic Acids Research,* 30**,** 866-75.

Willis, A., R., Torraca, V., Gomes Margarida, C., Shelley, J., Mazon-Moya, M., Filloux, A., Lo Celso, C. & Mostowy, S. 2018. *Shigella*-Induced Emergency Granulopoiesis Protects Zebrafish Larvae from Secondary Infection. *mBio,* 9**,** e00933-18.

Xian, W., Fu, J., Zhang, Q., Li, C., Zhao, Y.-B., Tang, Z., Yuan, Y., Wang, Y., Zhou, Y., Brzoic, P. S., Zheng, N., Ouyang, S., Luo, Z.-Q. & Liu, X. 2024. The *Shigella* kinase effector OspG modulates host ubiquitin signaling to escape septin-cage entrapment. *Nature Communications,* 15**,** 3890.

Xie, Z. & Tang, H. 2017. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. *Bioinformatics,* 33**,** 3340-3347.

Xu, D.-Q., Cisar, J. O., Ambulos Jr, N., Burr, D. H. & Kopecko, D. J. 2002. Molecular cloning and characterization of genes for *Shigella sonnei* form IO polysaccharide: proposed biosynthetic pathway and stable expression in a live *Salmonella* vaccine vector. *Infection and immunity,* 70**,** 4414-4423.

Xu, W., Zhang, Y., Huang, M., Yi, X., Gao, X., Zhang, D. & Ai, Q. 2015. The yesN gene encodes a carbohydrate utilization regulatory protein in *Lactobacillus plantarum*. *Annals of Microbiology,* 65**,** 115-120.

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. *Applied Environmental Microbiology,* 86.

Yang, F., Yang, J., Zhang, X., Chen, L., Jiang, Y., Yan, Y., Tang, X., Wang, J., Xiong, Z. & Dong, J. 2005. Genome dynamics and diversity of *Shigella* species, the etiologic agents of bacillary dysentery. *Nucleic Acids Research,* 33**,** 6445-6458.

Yang, J., Sangal, V., Jin, Q. & Yu, J. 2010. *Shigella* Genomes: a Tale of Convergent Evolution and Specialization through IS Expansion and Genome Reduction. In *Genomes of Foodborne and Waterborne Pathogens*, ASM Press

Yazyu, H., Shiota-Niiya, S., Shimamoto, T., Kanazawa, H., Futai, M. & Tsuchiya, T. 1984. Nucleotide sequence of the *melB* gene and characteristics of deduced amino acid sequence of the melibiose carrier in *Escherichia coli. Journal of Biological Chemistry,* 259**,** 4320-6.

Yilmaz, C., Rangarajan, A. A. & Schnetz, K. 2020. The transcription regulator and c-di-GMP phosphodiesterase PdeL represses motility in *Escherichia coli*. *Journal of Bacteriology,* 203.

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 Microbiology Letters,* 277**,** 197-204.

Zakikhany, K., Harrington, C. R., Nimtz, M., Hinton, J. C. & Römling, U. 2010. Unphosphorylated CsgD controls biofilm formation in *Salmonella enterica* serovar Typhimurium. *Molecular Microbiology,* 77**,** 771-86.

Zhang, X., Payne, M., Nguyen, T., Kaur, S. & Lan, R. 2021. Cluster-specific gene markers enhance *Shigella* and enteroinvasive *Escherichia coli* in silico serotyping. *Microbial Genomics,* 7.

Zhang, Z., Aboulwafa, M., Smith, M. H. & Saier, M. H., Jr. 2003. The ascorbate transporter of *Escherichia coli. Journal of Bacteriology,* 185**,** 2243-50.

Zimin, A. V. & Salzberg, S. L. 2020. The genome polishing tool POLCA makes fast and accurate corrections in genome assemblies. *PLOS Computational Biology,* 16**,** e1007981.

Zychlinsky, A., Prevost, M. C. & Sansonetti, P. J. 1992. *Shigella flexneri* induces apoptosis in infected macrophages. *Nature,* 358**,** 167-169.