THE GENETICS AND EPIDEMIOLOGY OF *SHIGELLA SONNEI* AND *SHIGELLA FLEXNERI* IN VIETNAM

Benjamin Robert Sobkowiak

A thesis submitted for the degree of Doctorate of Philosophy

Department of Genetics, Evolution and Environment

University College London

I, Benjamin Robert Sobkowiak, 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.

……

Abstract

Shigella sonnei is rapidly emerging as the primary agent of bacillary dysentery, or shigellosis, in many developing countries, replacing the historically more prevalent species, *S. flexneri,* in these regions. There have been various theories proposed to explain this phenomenon, including environmental changes and increased antimicrobial use, though the precise reasons for this shift are still uncertain. Here I present four studies investigating key ecological and genetic differences between *S. sonnei* and *S. flexneri* from a region that has undergone this pattern of species replacement, Vietnam. This work combines experimental and bioinformatics techniques with the aim of identifying the extent that differences in disinfectant sensitivity, chromosomal antimicrobial resistance profiles and gene content will contribute to the successful spread of *S. sonnei* over *S. flexneri.*

Firstly, I conducted *in vitro* experimental work to characterise differences between species with respect to resistance to chlorine disinfection and tolerance to the detergent SDS. The mechanisms by which the bacteria respond to this treatment, in particular the role of efflux pumps, were then explored to determine whether any informative variation in these systems will explain any differences in disinfectant sensitivity. The availability of high quality whole genome sequences for ~150 of each *Shigella* species allowed for robust bioinformatics work to describe genomic variation between species. These sequences are used to detect key resistance mutations in each species and look for associated fitness compensating mutations. Finally, the complete genome sequence of all coding regions in each strain was *de novo* assembled to look for species-level gene content variation that might contribute to functional differences between *S. sonnei* and *S. flexneri.* The results of these studies show that there are clear biological differences between *S. sonnei* and *S. flexneri*, though more work is necessary to fully elucidate the reasons for the species replacement in developing countries.

Acknowledgments

I am very thankful to everyone that has contributed to the completion of this thesis, both in the production of its content and the emotional support afforded to me throughout, by friends, family and colleagues.

Firstly, I would like to thank my Ph.D. supervisor Francois Balloux for giving me the opportunity to undertake this research and guiding me throughout the process by constantly challenging me to develop my ideas and to never lose sight of the biological significance of my work. The lessons I have learnt under his supervision have already enabled me to advance my scientific endeavours and no doubt will continue to do so throughout my career.

All members of the Balloux group, past and present, have helped in many ways to develop this work. In particular, I would like to thank Adrien Rieux, Matteo Fumagalli, Matteo Spagnoletti, Stephen Price, Liam Shaw and Florent Lassalle for their academic and personal support throughout this project. I would also like to thank Vegard Eldholm for teaching me so much in a short time about microbial pathogens and microbiology. In addition, I am grateful to Stephen Baker for imparting his knowledge of bacterial pathogens and always being quick to share data and ideas. I also thank my current supervisor Taane Clark for providing me the time to finish this work.

I thank the Natural Environmental Research Council - NERC and the Department of Genetics, Evolution and Environment at UCL for funding this work.

On a personal level, all my friends and family that have provided support and distraction since I began this degree have been extremely important in allowing me to complete this research. Notably, my close family, friends and girlfriend that have all helped to inspire and keep me going through the hardest times.

Finally, I would like to dedicate this thesis to my mum, Deborah Fielder, who has always done everything to enable me accomplish my goals. I appreciate everything that she has done and this would not have been possible without her.

Thank you all. Ben

Table of contents

Chapter 1

General Introduction

1.1 Bacterial Pathogens

Infectious disease epidemiology

Bacteria are among the most diverse and abundant organisms on Earth, with the sum of bacteria living on a single human individual outnumbering human cells 10 to 1[1]. Many bacterial species are human pathogens, with a number of important diseases caused by microbial agents, including tuberculosis, meningitis and pneumonia. As a consequence, bacterial infections continue to cause a great number of deaths every year in humans. Discovering treatments for these diseases and preventing their transmission is of the upmost importance for global health.

At the beginning of the 20th century many bacterial diseases were almost untreatable, contributing to low life expectancies and high infant mortality rates[2]. With the success of antibiotics around the mid-century, as well as a general improvement in human health and the scientific understanding of the biology of these organisms, the threat from such infections was declining in the developed world, though they remained a cause for concern in developing countries, notably with diarrheal diseases in children[2]. The end of the 20th century brought global cases of newly emerging microbial infections and the re-emergence of many diseases that were thought to have been eradicated in the developed world, such as tuberculosis, cholera and infections from *Streptococcus aureus*[3,4]. The awareness that infectious diseases have been increasing globally and understanding the reasons for the rise in the number of cases has led to an emphasis on research to address this growing concern. Major threats to human populations also include viral and fungal diseases such as human Immunodeficiency virus (HIV) and *Aspergillosis*, though for the purposes of this report I will be concentrating only on the transmission of bacterial pathogens. The main factors that influence the present-day transmission of bacterial pathogens are shown in **Figure 1**.

A rise in human migration and recreational travel has changed the accessibility of susceptible individuals to certain diseases, as the chance increases of transmission of novel bacterial infections that would not have been able to pass between previously disparate groups. There can also be different genetic and social risk factors associated with disease transmission that may be influencing the prevalence of particular infections in a more connected and diverse populace[5].

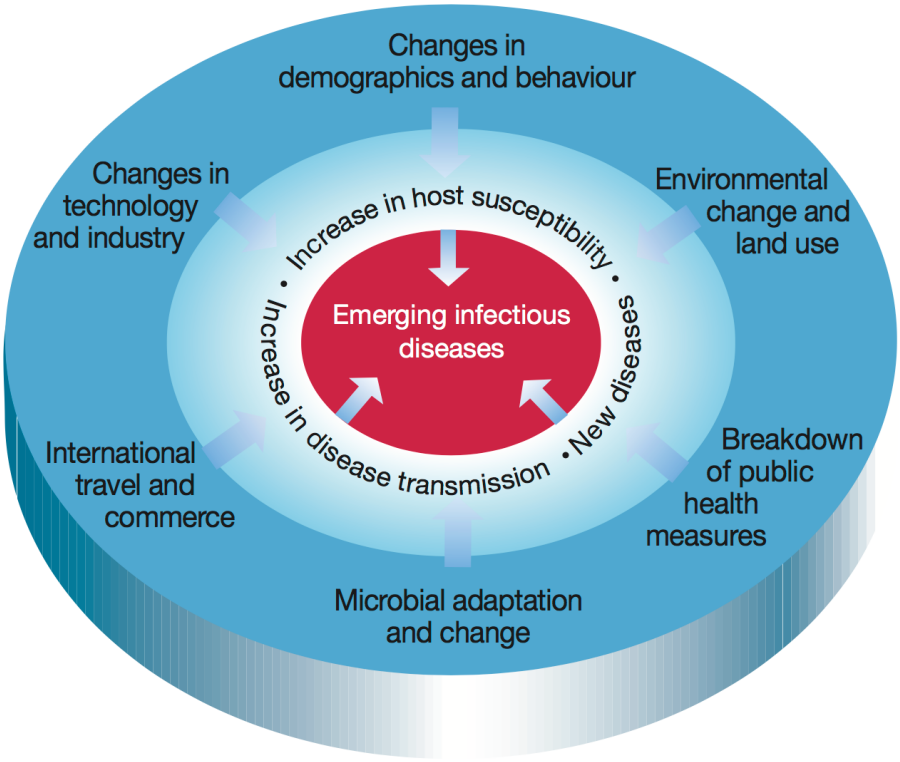


Figure 1. Factors influencing the emergence and spread of human infectious bacterial diseases. Illustration taken from Cohen[2].

Human population demographics, notably in developed nations, are transforming due to societal and behavioural changes. Many of these countries have an aging population as people wait longer to have children and as advances in healthcare leads to higher average life expectancies. This can increase the number of individuals who are susceptible to particular illnesses through weakening immune systems with age or can increase the transmission and reactivation of chronic or latent infections as individuals have a greater chance of harbouring a disease for a longer time[2,6].

The risk of bacterial disease transmission has also increased through exposure to particular infections with changes in human behaviour, such as the increase in people living in close proximity in densely populated cities, dietary changes and changes in sexual behaviours. The increases in HIV both in men who have sex with men (MSM) and heterosexual populations has been linked to high risk sexual behaviours[7,8]. The rate of these behaviours has increased in the developed world from the mid 20th century with the sexual revolution, and through greater economic disparity in many less developed regions, such as sub-Saharan Africa, where there are reports of higher numbers of sexual partners due to

increases in the numbers of sex workers and through pronouncement of unsafe social behaviours, with women particularly at risk[9]. There is evidence for co-incidence of HIV and many bacterial diseases, for instance tuberculosis, where HIV positive patients are more likely to contract the disease with a faster progression of symptoms[10]. An increase in sexual risk behaviours can also directly increase the frequency of sexually-transmitted bacterial infections such as gonorrhoea and chlamydia, which can cause severe disease and infertility if not adequately treated.

Dietary changes have also had an impact on bacterial disease transmission, with a revolution in eating habits and food preparation habits in many countries. Particularly in the developed world, people will eat a wider range of food from non-local sources that may be more likely to carry novel pathogens from their point of origin[11]. There is also a greater number of people eating outside of the home, and this has been linked to an increase in the transmission of food-borne pathogens[12]. Additionally, it has long been known that there is an association between the risk of contracting zoonotic diseases and agriculture through close contact between host and pathogen. The intensification of farming practices to feed a growing populace and a shortage of arable land in many areas due to climate change and overpopulation has led to large-scale farms of genetically homogenous livestock where disease can spread rapidly, with an increased risk of transmission into other taxa. This in turn has increased the frequency of zoonosis emergence and re-emergence in both human and animal populations[13,14].

In addition to contributing to an increase in zoonotic diseases described above, change to the natural environment and climate has contributed to the emergence or increased incidence of many other human pathogens[15]. Warmer climates and the rising of annual rainfall and sea temperatures have all been linked to an increase in a number of bacterial infections, particularly diarrheal and food-borne diseases[16]. As an example, *Vibrio* diseases from contaminated water or seafood have been shown to be increasing worldwide in both humans and aquatic mammals through rising sea surface temperatures (SST). Rising SST causes changes to bacterial growth as well as altering interactions with other organisms that can act as a vehicle for transmission to humans, such as oysters[17]. In addition, climate change can lead to increased reproduction in some bacterial pathogens, such as some salmonellas, and changes to microflora composition as average temperatures rise[18]. Forecasting the diseases that will emerge in particular regions and the risk posed to human populations depends on a multitude of factors including the optimum temperature range and adaptability of a pathogen[15]. Modelling potential changes to bacterial species compositions

in the face of a changing natural environment will be vital for the effective management of infectious diseases.

Finally, advances in technology and medical care have been responsible for increasing the chance of host survival with both infectious and chronic diseases, though these treatments can leave people more vulnerable to further infection. A combination of lengthier stays in hospital, extended and varied courses of treatment and potential inadequacies in controlling contamination have created environments where pathogens can be transmitted readily between susceptible individuals, as well as leading to longer infections that allow time for microbial adaptations for increased virulence. There is evidence of severe outbreaks of bacterial infections that have undergone mutation to alter their pathogenicity in these hospital settings, famously with the evolution of Methicillin-resistant *Staphylococcus aureus* (MRSA)[19]. Advancement in the treatment of infectious disease has also heralded the dawn of the era of widespread antibiotic use and subsequent acquisition of antimicrobial resistance in bacteria. This has had one of the most profound effects on pathogen transmission and this subject will be discussed in detail in section 1.2.

Mechanisms of pathogenicity

Bacterial pathogens have evolved variety of mechanisms to evade host detection or elicit a cellular response to increase their virulence **(Figure 2)**. These strategies will damage the host directly or help the organism to battle against a host immune response, with many bacteria using a combination of these mechanisms to cause disease. Species such as *Streptococcus pneumoniae* can produce large polysaccharide capsules that surround the bacteria and provides protection against phagocytosis by preventing the host immune response from recognising the pathogen antibodies as a foreign invader[20]. The cell wall of both Gram-negative and Gram-positive bacteria contains components (lipopolysaccharides, LPS, and teichoic acid respectively) that are toxic to the host and can cause septic shock when released upon cell death and lysis. Bacteria can also secrete four categories of exotoxins, (1) two subunit A-B toxins, (2) proteolytic toxins that break down host proteins, (3) membrane-disrupting pore forming toxins, and (4) other toxins that can modify the host structure or disrupt signalling pathways[21,22]. Shiga and Shiga-like toxins found in *Shigella* species are examples of the two sub-unit A-B toxins[23].

The ability to avoid host clearance by adhering to surfaces is also an important mechanism to increase virulence. A host will employ a number of mechanical strategies to clear infections such as coughing, sneezing and increasing blood flow, and some bacteria

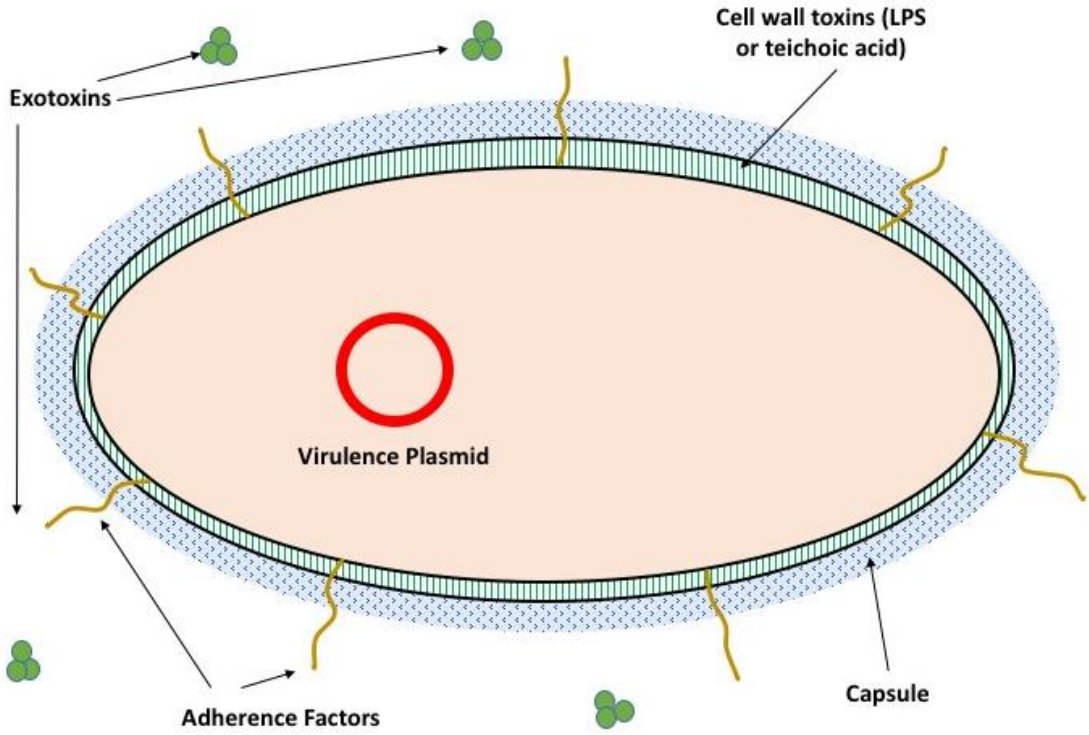


Figure 2. An illustration showing some common bacterial virulence factors.

produce either fibrous and non-fibrous protein (fimbriae) or polysaccharide adherence factors to attach themselves to host surfaces including skin, the mucosal lining of the nasopharynx, urogenital area and oral cavity, and deeper tissues such as the gut epithelia[21]. Not only does this adherence enable the pathogen to evade host clearance, but this close contact allows the pathogen to invade the host and continue the infectious cycle[21]. Some bacteria, including *Streptococcus aureus,* will access the host through extracellular invasion where enzymes that degrade host tissues are released and allow the pathogen to infect the host whilst remaining external to the cell[24]. Intracellular invasion allows the pathogen to live within the host cell and co-opt host resources for survival and growth[25]. Many bacterial species can live both within and outside of cells (facultative intracellular parasites), including several *Mycobacterium* species and *Salmonella typhi,* whilst others (obligate intracellular parasites) survive exclusively in an intracellular lifestyle, such as *Chlamydia* species and *Mycobacterium leprae*[26]*.*

Understanding the mechanisms by which microbes can evade host detection and cause disease is important for developing effective treatments. Evidence has shown that the loss and gain of genes has been integral in the evolution of pathogenesis in bacteria[27], and with the horizontal gene transfer (HGT) of complete genes and large genetic elements, bacterial

strains are able to rapidly acquire new virulence plasmids and mechanisms to compete against host defences. The ongoing battle between host and pathogen to develop antagonistic strategies, along with the ubiquity of HGT for transferring large amounts of genetic information readily between microbial populations, has brought about perhaps the greatest threat that bacterial pathogens pose to human populations in modern times; the acquisition of antibiotic resistance.

1.2 Antibiotic Resistance and Compensatory Mutation

The use of antibiotics as an effective treatment for bacterial disease has markedly improved human health for the past century and rendered some infections that were almost always fatal, such as meningitis, treatable in all but the most serious cases[28]. Unfortunately, the misuse and often overuse of these drugs has led to a reduction in their efficacy due to the increased frequency of bacteria evolving antibiotic resistance. Coupled with the fact that few novel antibiotics are currently in development[29], a future where previously treatable infections such as typhoid and cholera are major global concerns is a very real possibility. Understanding the evolutionary principles underlying the acquisition and spread of resistance is one of the most important challenges facing scientists today.

Numerous studies have shown that antibiotic-resistant bacteria were present at low levels before the introduction of antimicrobial drugs, with evidence of naturally occurring resistance conferring competitive advantage over antibiotic producing bacteria[30]. It is clear though that the increased selection pressure of widespread antibiotic drug therapy has accelerated the spread of resistance genes amongst bacterial populations through adaptive evolution.

A growing concern in recent times is the emergence of multidrug resistance (MDR) in bacteria, where an organism has evolved resistance to more than one antibiotic. Often, in the simplest cases of resistance, a bacterial pathogen will acquire resistance to first-line antibiotics, typically one that is the primary treatment for a given infection as it has the best risk-benefit profile when taking into account the cost of production and safety of the drug. If a patient carries an infection that does not respond adequately to these first line drugs, often due to resistance in the bacteria, further courses of antibiotics that have a lesser risk-benefit profile may be administered, referred to as second- and third-line drugs. Given the adaptive nature of the evolution of resistance[31], the increased use of these second- and third-line drugs in cases where the infectious agent is first-line antibiotic resistant can lead to

sequential resistance to these antibiotics over time, and thus the pathogen will be considered MDR.

Last-line antibiotics are so called as they are often the drugs of last resort and administered only when all other therapies have been ineffective or when only one antibiotic treatment exists for a particular pathogen. Clinicians are often averse to issuing these treatments due to the rate at which bacteria are acquiring resistance; indeed, there are now cases where pathogens have evolved resistance to drugs classified as last-line, such as recent evidence plasmid-mediated colistin resistance in *Escherichia coli* from China[32]. Pharmaceutical companies have become reluctant to fund the development of new antibiotics for a number of reasons including low profitability, in part owing to numerous cases of bacteria acquiring resistance, and this has been a major factor in the paucity of new antibiotics entering the market. As a result, there have been no new antibiotics for over 20 years and, more worryingly, the last class of antibiotics developed to treat Gram-negative bacteria, fluoroquinolones, coming more than 40 years ago[33].

It has become increasingly important for evolutionary biologists and microbiologists to research the mechanisms by which antibiotic resistance can be acquired and fixed in populations. Resistance genes occur either through spontaneous *de novo* mutation within the chromosomal genome or by HGT from other bacterial species **(Figure 3)**. The most common resistance mutations will correspond to alterations in the target site of the antibiotic, increases in drug efflux, and gene amplification[34].

The rate at which these genes evolve and spread is influenced by the interaction between the selective force exerted by the antibiotic, the rate of supply of these mutations and the fitness effects of carrying the resistance gene, both in the presence and absence of the antibiotic[31]. Evidence has shown that number of chromosomal mutations conferring antibiotic resistance will have a deleterious effect on fitness, with reduced bacterial growth typically observed[35] as loci where mutations occur may also be involved in vital metabolic functions[36], and with these true *de novo* mutations a population has not passed enough time to evolve strategies to reduce these costs. To a lesser extent there can be also be costs associated with harbouring plasmid-mediated resistance through the acquisition and maintenance of these genetic elements. These costs are often lower though as often an organism can already carry adaptations to account for plasmid carriage[37] or, as the resistance has evolved previously in another host, there may already have been selection acting on this genetic element to mitigate associated fitness costs[38]. Importantly though, plasmids can carry a number of simultaneous resistance genes to a variety of antimicrobial

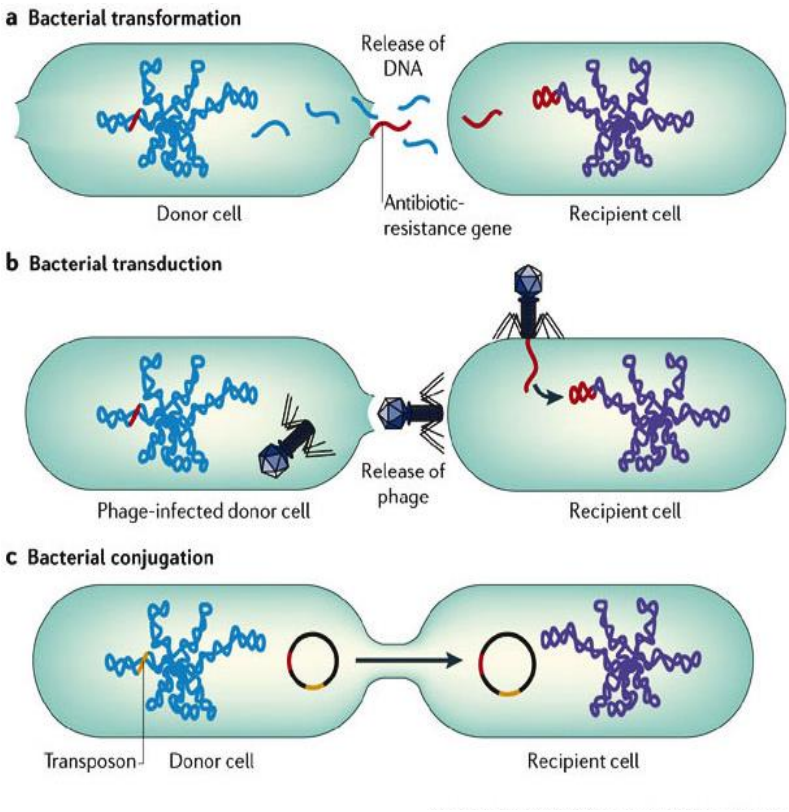


Figure 3. Illustration of three mechanisms leading to horizontal gene transfer (HGT) in bacteria. (a) Bacterial transformation occurs when DNA is released through cell lysis of the donor cell and taken up by the recipient, where it can then be integrated into a chromosome or existing plasmid. (b) Bacterial transduction is when a recipient cell acquires genetic information from a donor cell through a bacteriophage vector, also called lysogeny. (c) Bacterial conjugation is the sharing of genetic information, notably plasmids and transposons, through direct contact or a mating bridge between two bacterial cells. Illustration taken from Furuya and Lowry[39].

families and it has been shown that there are higher fitness costs associated with the acquisition of plasmids harbouring more resistance genes, though not with increasing plasmid size, suggesting that resistance genes on these elements carry a greater cost than other genetic elements[38].

Theoretical models suggest that, in cases of increasing antibiotic resistance in populations, reducing antibiotic use would enable susceptible strains to outcompete the resistant strains as the selective advantage of resistance would be diminished and, over time, the deleterious resistance gene will be lost from the population, or be reversed back to wild-type[40]. More recently it has been shown though that this theory may be complicated by the evolution of compensatory mutations, with evidence suggesting that these mutations are more likely to occur than reversions[36]. These genes, whilst deleterious on their own, can increase the fitness of resistant bacteria to that of susceptible strains resulting in stabilization of resistance genes in the population. This can also increase the rate of resistance genes

spreading through the population when antibiotics are present as organisms with both resistant and compensatory mutations will be at a greater selective advantage over susceptible strains without any of the associated fitness costs[41].

There have been several examples where compensatory mutations have been identified in pathogens, such as with streptomycin resistance in a number of organisms including *Mycobacterium tuberculosis* and *Escherichia coli*[41]. Mutations in the gene encoding for the ribosomal protein S12 (*rpsL)* can confer resistance to streptomycin but also greatly reduces the efficacy of translation. *In vitro* studies have shown that when *rpsL* mutants are grown in streptomycin-free cultures the resistance genes remain in the population and secondary mutations in ribosomal proteins S4, S5 and L19 can occur[42]. These mutations are postulated to compensate the associated costs of resistance, and the relative fitness of mutant populations will return to close to that of wild-type strains.

The mechanisms by which compensatory mutations reduce the fitness costs associated with resistance genes can be somewhat varied. Perhaps the most common is the restoration of a lost or compromised function by intragenic and intergenic mutations. Intragenic compensations within the same gene as the deleterious mutation will restore the read frame of the original gene where the resistance mutation is either an insertion or deletion. These have been identified in bacteria resistant to various antibiotics including streptomycin, β-lactams and rifampicin[36]. Compensatory mutations can also be in the form of intergenic mutations in other areas of the genome to the resistance gene such as with streptomycin resistance in *E. coli* described previously[42].

A further compensatory mechanism is to bypass the altered genomic function by replacing it with an alternative pathway. This has been demonstrated in isoniazid-resistant *M. tuberculosis* with a mutation in the *katG* gene[43]. This mutation will also incur a fitness cost to the organism by reducing the activity of the catalase-peroxide enzyme, important in the oxidative stress response. Secondary mutations in the *ahpC* gene will stimulate the overproduction of alkyl hydroperoxidase, which will compensate for the reduction in KatG catalase. Finally, where the cost to fitness is due to a defect in a transcribed enzyme, secondary mutations that promote the overproduction of this enzyme itself can compensate for the reduced function at uncompensated levels. In many cases though there is a complex interaction involving multiple mechanisms, and there is also evidence of epistatic fitness effects with different combinations of resistance and compensatory genes[44,45].

1.3 The Genus *Shigella*

History and classification

The first description of the pathogen responsible for causing bacillary dysentery was in 1898 by Kiyoshi Shiga upon the discovery of *Bacillus dysenteriae,* later renamed *Shigella dysenteriae,* from a severe outbreak in Japan[46]. In the 1950's this species was placed into a new genus, eponymously named *Shigella*, compromising four species based upon serological and biochemical characteristics; *S. dysenteriae, S. boydii, S. flexneri* and *S. sonnei.* These species, or serogroups, are further subdivided into serotypes based on the O antigen of a lipopolysaccharide outer membrane of the cell wall, serogroup A (*S. dysenteriae,* 12 serotypes), serogroup B (*S. flexneri,* 6 serotypes), serogroup C (*S. boydii*, 16 serotypes), and serogroup D (*S. sonnei*, 1 serotype)[47]. *S. flexneri* and *S. sonnei* are the two most common species isolated globally, whilst *S. boydii* appears to be restricted to the Indian sub-continent[47]; *S. dysenteriae* is rarely found today but has historically been responsible for severe outbreaks in developing regions[48–50]. Of these groups, serogroup D, *S. sonnei*, is the most physiologically distinct and can be distinguished by key biochemical and genetic characteristics, including lactose fermentation and differences in the utilization of a range of substrates[51].

Pathology and virulence

Shigella species are pathogenic Gram-negative bacteria that can cause bacillary dysentery, or shigellosis, in humans. Clinical symptoms of *Shigella* infections can range from mild diarrhoea to very severe, bloody dysentery and can cause death in some cases, mainly in young children. They are highly transmissible pathogens, with one study with human volunteers finding that some strains of *Shigella* can cause active infections with exposure to as few as 10 bacteria[52]. Infections are spread through the faecal-oral route though there is evidence of environmental and sexual transmission, including high incidences of infection in HIV positive patients[49]. The most severe infections are due to the inflammation of the gut epithelia leading to the destruction of the colonic mucosa[53]. The mechanism of epithelial entry is controlled by two gene loci on the pINV invasion plasmid carried by all *Shigella* strains; the type III secretion system (TTSS) *mxi-spa* locus[47] and the *ipa* locus that encodes three genes for actin polymerization at the site of entry and cytoskeleton rearrangement modulation[53]. The disease is treated with a variety of antibiotics that varies depending on the serogroup and availability of drugs in the infected population, though many previously common therapies such as ampicillin, sulphonamides and fluoroquinolones are now becoming less efficient due to the ability of the pathogen to rapidly gain antimicrobial resistance both through chromosomal mutations and plasmid acquisition by HGT[49].

Evolutionary origin

The evolutionary relationship between *Shigella* and *Escherichia coli* and the designation of *Shigella* as a separate taxonomic group is contentious, with *Shigella* falling as a paraphyletic group or as monophyletic subclades within *E. coli* in a number of studies[23,47,51,54,55]. Prior to the use of genetics to compare organisms, *Shigella* species were known to be very closely related to *E. coli* based on shared physiological characteristics but were originally assigned a distinct classification as they were pathogenic and able to survive in a wide variety of environments, whereas *E. coli* was regarded as a relatively harmless commensal bacteria[55]. With the discovery of pathogenic *E. coli* in 1944, now known as Enteroinvasive *E. coli* (EIEC), and later characterisation of another five strains of disease causing *E. coli,* of which one includes the common strain O157:H7 that also produces a Shiga-like toxin found in *S. dysenteriae*[56], this consensus had to be revised. With advances in molecular genetics techniques for discerning associations between species many studies sought to determine the true phylogenetic relationship between *Shigella* and *E. coli,* and reconstruct the evolutionary origin of *Shigella* serogroups.

Initially, studies looking at the genetic divergence through DNA re-association between strains of *Shigella* and *E. coli* found high levels of reciprocal binding between the species and advocated that *Shigella* species be reassigned to the *Escherichia* genus[57,58]. More sophisticated techniques followed that compared the sequence similarity of chromosomal housekeeping genes[23,51] and orthologous sequences[59] to reconstruct the phylogenetic tree of *E. coli/Shigella,* and found that *Shigella* serogroups were positioned in clusters within *E. coli*[23] or formed a single pathovar of *E. coli* with EIEC[51,60]. These results cast doubt on the legitimacy of *Shigella* as a separate species and instead advocated a close evolutionary relationship with *E. coli* and the designation of *Shigella* strains as specialized *E. coli* that have acquired virulence independently in multiple lineages [55].

The availability of whole genome sequencing (WGS) allowed for extensions on these methods to include all shared (core) genes rather than a subset of housekeeping genes. Again, when reconstructing the evolutionary relationship between *Shigella* species and a range of pathogenic and commensal *E. coli* with these core alignments, *Shigella* were placed within *E. coli,* with *S. dysenteriae* in particular clustering within the pathogenic *E. coli*[61,62] **(Figure 4A)**. Furthering the application of whole genome data, work by Sims and Kim[55] sought to resolve the *E. coli/Shigella* relationship by comparing strains using an alignment-free feature frequency profile (FFP) to consider all genomic features not limited to core genes. They found conflicting results when comparing the trees obtained through all possible features and filtered for only core, shared features. In the all feature phylogeny

Shigella formed a single monophyletic clade whereas when only core features were used there were two related subclades, with *S. dysenteriae* again being placed in a separate cluster outside of the other *Shigella* species[55]. In contrast to the indication of a multiple origin hypothesis for *Shigella* evolution, a study by Zuo *et al*[63] using a WGS gene protein product approach (CVTree) to compare the genomes of *Shigella* and *E. coli* found evidence for a single *Shigella* clade, suggesting that rather than *Shigella* falling within the *E. coli* species, it could be considered a highly related group within the *Escherichia* genus that has evolved from a single ancestor **(Figure 4B)**.

The history of the major invasion plasmid in *Shigella,* pINV, a pivotal step in the evolution of pathogenesis, is also not fully resolved, with studies suggesting both a single acquisition in an *E. coli* ancestor prior to speciation and multiple convergent events across *Shigella* lineages. Work by Lan *et al*[64] supports the multiple origin hypothesis of *Shigella* serogroups and the repeated acquisition of the virulence plasmid, including in the EIEC, of which there are in fact two forms, pINV A and pINV B, that contain largely homologous genes but with distinct differences. Escobar-Paramo *et al*[54] did not detect these different forms of the plasmid though when comparing three virulence genes (*ipaB, ipaD* and *icsA)* in strains of *Shigella* and EIEC, supporting the model of a single origin of pINV in these groups from a single *E. coli* strain. A different hypothesis was put forward by Yang *et al*[65] who suggested that the virulence plasmid was obtained at multiple times across *Shigella/*EIEC lineages from a diverse ancestral pool of plasmids rather than just the two forms postulated by Lan *et a*[64]*.*

Clearly the exact relationship between *Shigella* and *E. coli* and the evolution of virulence in *Shigella* is still uncertain, with evidence both for the single event radiation of *Shigella/*EIEC from a single ancestor as well as a multiple origin hypothesis with subsequent convergent evolution of virulence among *Shigella/*EIEC lineages. Similarly, the designation of *Shigella* as a distinct species from *E. coli,* particularly from EIEC, is contentious and many studies now support the classification of *Shigella* as a separate genus solely for historical reasons; though with advances in molecular genetic techniques and the availability of more whole genome sequences this matter may yet be resolved.

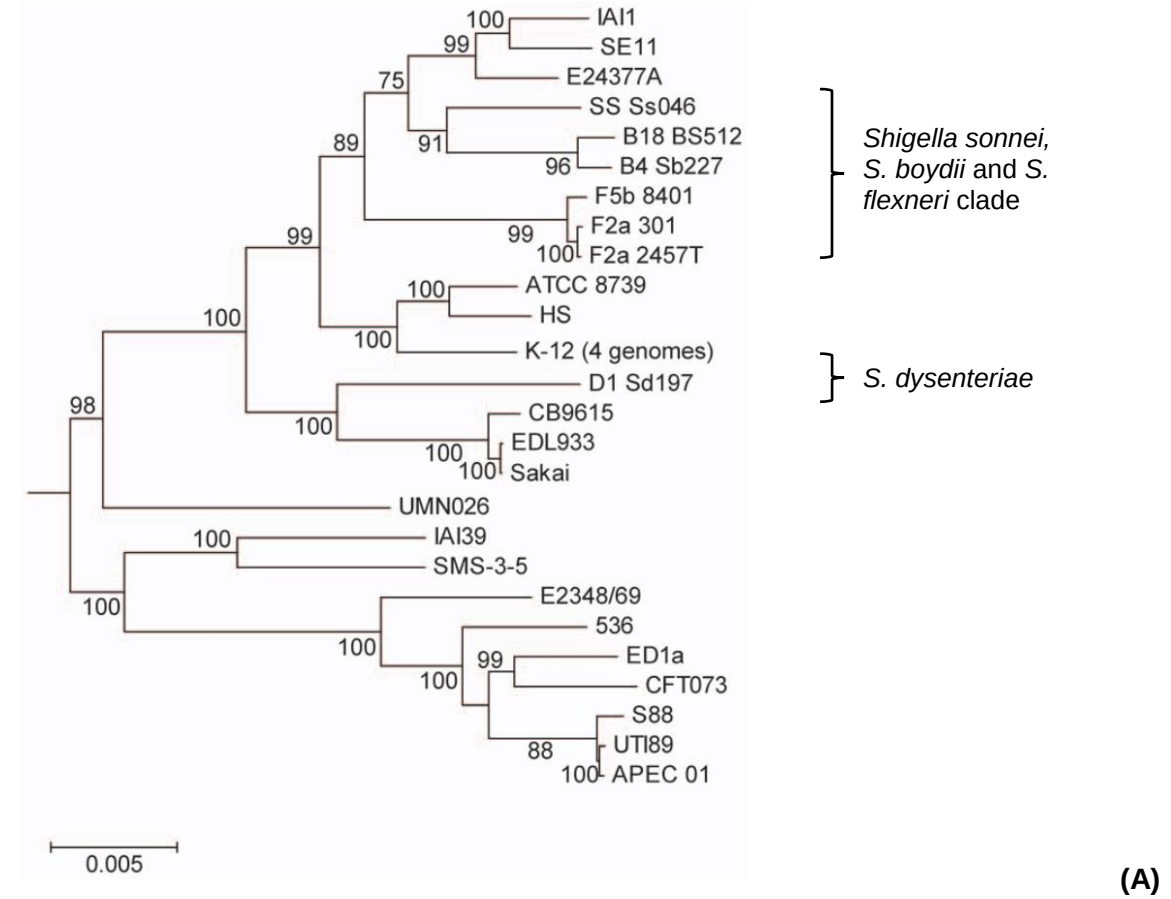


Figure 4. Conflicting phylogenetic trees reconstructing the evolutionary relationship between *Shigella* species and *E. coli.* (A) Adapted from Zhou *et al*[62], a maximum likelihood tree of 19 *E. coli* and 7 *Shigella* strains built from 2,034 core genes, showing *Shigella* species to form two clades within *E. coli.* **(B)** From Zuo *et al*[63], CVTree approach to construct trees based on whole genome, alignment free comparison of protein sequences of *Shigella* species and *E. coli.* K = length of peptide chains. This approach finds *Shigella* and *E. coli* to be separate, sister clades.

1.4 *Shigella* in Vietnam

Shigellosis is a major public health concern in much of Asia, with the number of annual infections estimated at 125 million in this region alone, resulting in 14,000 deaths[66]. Historically, the most commonly isolated species from infections in Asia, including Vietnam, was *S. flexneri,* with this species the predominant agent of shigellosis in many developing countries. The less genetically diverse *S. sonnei,* whilst also found globally, was primarily isolated in more developed countries, for example Israel and USA, where it accounts for over 70% of all clinical presentations of bacillary dysentary[67]. There is evidence that the dominant serotype of *S. flexneri* can fluctuate over time and geography due to a changing temporal environments[68]. More recently though, there have been documented accounts of *S. sonnei* infections increasing in relative proportion in many countries where *S. flexneri* was previously the most common agent, such as Vietnam[67], China[69] and Brazil[70], with a simultaneous decrease in clinical presentations of *S. flexneri*. The reason for this shift in dominance is unclear, though there is a clear correlation with rapid development in these regions, coupled with improving sanitation and antimicrobial use, discussed further in section 1.5. Identifying the most commonly occurring serotype of *Shigella* is particularly important for the effective treatment and management of the spread of the disease as there can be variation in antibiotic resistance profiles and virulence between serotypes.

Vinh *et al*[67] first described the shifting pattern of species dominance in southern Vietnam after work by Seidlein *et al*[71] suggested that this pattern was being observed in other countries of Asia. They found that over a 14-year period there was a marked shift in the dominant species (from *S. flexneri* to *S. sonnei*) found in clinical isolates from patients with dysentery, along with a change in resistance profiles of the strains that were circulating **(Figure 5)**. Such was the dramatic increase in the number of *S. sonnei* infections in this study period that the proportion rose from 29% to 78% of total shigellosis presentations. Also within this period, there was an overall increase in the number of strains showing resistance to a range of commonly used antimicrobials, with over 80% resistant to three or more of the seven tested antibiotics towards the end of the study compared to 63% in the first period of study. Resistance to nalidixic acid increased significantly in this time period and though rarely recommended for use as an antimicrobial directly, resistance to this compound is the same mechanism conferring resistance to fluoroquinolones, which are often one of the first line treatments of shigellosis. Interestingly, the severity of the disease was also greater in the last period of the study, with an increased number of patients reporting watery diarrhoea, pains and convulsion. These are symptoms associated with severe infections and, conversely, are more likely brought on from *S. flexneri* or *S. dysenteriae* than *S. sonnei*.

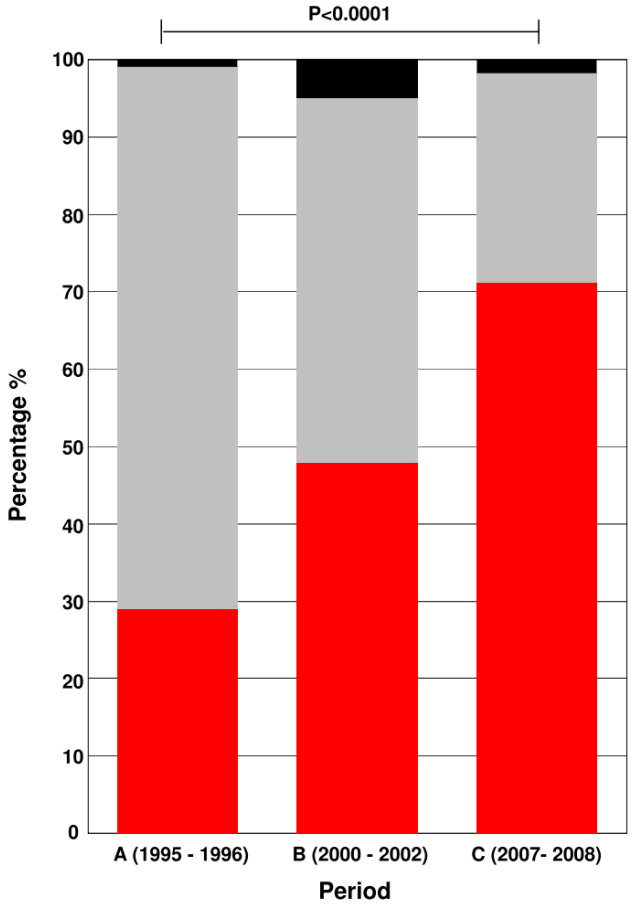


Figure 5. The relative proportion of *Shigella* species in Vietnam between 1995 and 2008. Data were collected on childhood Infections over three periods in southern Vietnam. *Shigella sonnei* are coloured red, *Shigella flexneri* grey and other *Shigella* species in black.

Following this report detailing the shifting pattern of shigellosis infections in Vietnam, a study by Holt *et al*[72] then looked to reconstruct the phylogenetic and epidemiological history of *S. sonnei* in the region to determine key events that may have led to the expansion of the organism. Using whole genome sequencing of clinical strains of *S. sonnei* over 15 years, the team were able to reconstruct a likely most common recent ancestor (MCRA) within the 1980's that had been introduced to Ho Chi Minh City from Europe. The subsequent spread of the pathogen through the region is proposed to have likely been due to clonal expansion through four selective sweeps that have coincided with decreased genetic variation within the population but with an increase in virulence and antibiotic resistance. This suggests that these bottlenecks in genetic diversity were driven by increased selective pressure on the population and the resultant fixation of key mutations from more successful *S. sonnei* clonal sub-populations. The data collected in this study also indicate that by 2010 the proportion of shigellosis infections caused by *S. sonnei* is up to 99% in southern Vietnam.

The majority of the strains sequenced in this study were found to be resistant to more than one antimicrobial and it has been hypothesized that the MCRA was already multidrug resistant (MDR) at its introduction. As with the Vinh *et al*[67] report, resistance to fluoroquinolones was very high within the population along with resistance to a number of other antimicrobials, either through chromosomal mutations or plasmid acquisition. In particular, plasmid acquisition appears to be an important driver in the establishment of *S. sonnei* in the region, with the horizontal acquisition of the plasmid-borne colicin system from other enteric bacteria facilitating the inflammation within the human gut in *S. sonnei* infections and allowing the organisms to out-compete other microbiota. The fixation of this plasmid also coincided with the first genetic bottleneck in the population. In addition, resistance to third-generation cephalosporins (plasmid-mediated) and fluoroquinolones (chromosomal) was fixed in subsequent bottlenecks within a short time[73], demonstrating the rate at which bacterial populations can acquire multiple resistance mechanisms that may confer a competitive advantage and accelerate the spread of a pathogen.

1.5 The Shifting Dominance of *Shigella* Species in Developing Countries

The pattern of species replacement observed in Vietnam and many developing countries suggests there may be important genetic and epidemiological distinctions between *S. sonnei* and *S. flexneri* that influence each species' survival and transmission in different environments. As stated previously, *S. sonnei* has historically been the most common cause of shigellosis in more developed regions, such as the USA[74–76], Europe[77–79] and Israeli cities[80], and the increase in incidence in lower-income countries has corresponded with an improvement in water quality and sanitation, as well as antimicrobial use. This would suggest that *S. sonnei* strains emerging in these regions have a competitive advantage over *S. flexneri*. It may be that the genetic and phenotypic features of the strains that are entering these regions are more optimally suited to surviving in these changing conditions, such as already MDR *S. sonnei* emerging in Vietnam, or that *S. sonnei* are characteristically better at adapting to any change in the environment, such as a greater ability to take up antibiotic resistance genes from other bacteria.

Indeed, it appears from the previous work conducted by in Vietnam, an increase in antibiotic resistance is a key feature in *Shigella* populations in this region, and in particular in the emerging clonal strains of *S. sonnei*. A paper by Holt *et al* describes the emergence of *S. sonnei* infections globally, with all contemporary infections caused by a few clonal lineages that have expanded globally[81]. Of these, lineage III is the most abundant in Vietnam and the

rest of Asia[69,81,82], and these strains harbour plasmid-borne resistance to a wide range of antibiotics including streptomycin, tetracycline and sulphonamides[72]. In addition, localized adaptation since the clonal *S. sonnei* expansion within various locations in Vietnam brought about distinct CTX-M plasmids encoding extended-spectrum beta-lactamases (ESBLs) that confer resistance to a range of antibiotics including penicillin and cephalosporins, demonstrating ongoing recombination in *S. sonnei* in the region. This evidence that *S. sonnei* are able to readily acquire new resistance genes through HGT from other enteric and environmental bacteria suggests this species is able to acquire advantageous adaptive traits in response to selective pressures. Furthermore, there is evidence that increased levels of gut inflammation can facilitate conjugation and transfer of plasmid elements between pathogenic and commensal bacteria within the host. *S. flexneri* has been shown to inhibit the inflammatory response during infection in order to avoid the host detection and this may reduce the efficacy of HGT in *S. flexneri*, though inflammatory differences between *S. sonnei* and *S. flexneri* infections have not yet been investigated.

The frequency of mutations conferring resistance in *Shigella* species also appears to vary between species, both in de novo point mutations and the integration of genomic elements and plasmids. Work characterizing the resistance profiles of MDR *S. sonnei* and *S. flexneri* in Chilean children found some differences between the distribution of specific genetic determinants of both trimethoprim and ampicillin resistance in each species[83], with *S. sonnei* more frequently possessing the bla_{TEM} gene for ampicillin resistance on a conjugative plasmid while the blo_{OXA} chromosomal gene was the most prevalent in *S. flexneri*. This chromosomal ampicillin-resistance gene was also found in Japanese *S. flexneri* isolates, as well as a varying range of integron gene cassettes between species that conferred resistance to an assortment of antibiotics[82]. In addition, there have been further studies that have found differences in the range of resistance profiles and genetic determinants between local *Shigella* populations in a number of countries[68,71,84] **(Figure 6)**. These disparities may reflect differences in the evolutionary history of the populations, though there is evidence that *S. sonnei* appears to readily acquire a wide range of resistance to antimicrobials through plasmid conjugation that are not found in *S. flexneri.* At the current time there have been no studies to determine whether these plasmids can be transferred between *Shigella* species. Differences in antibiotic resistance profiles and the ability to acquire new genes when subject to selective pressures may present a competitive advantage for *S. sonnei* and contribute to the global spread of this species due to the widespread use of antimicrobials.

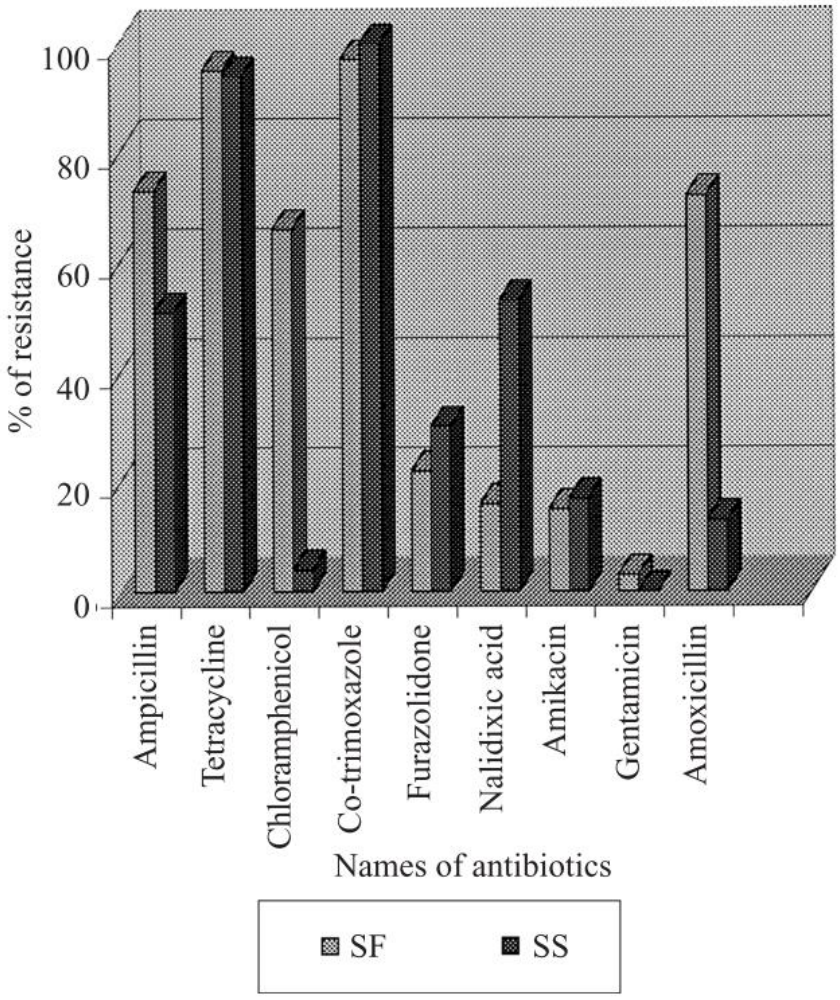


Figure 6. Antibiotic resistance in *S. sonnei* and *S. flexneri* infections isolated in Kolkata, India, 1995-2000. Graph taken from Dutta *et al*[68].

The increasing use of antibiotics and accompanying resistance is likely to be major feature influencing the emergence of *S. sonnei* as the primary cause of shigellosis in industrializing regions, though there are additional factors that have been postulated as contributing to the spread of *S. sonnei* and the replacement of *S. flexneri* in developing countries. Firstly, there may be a natural host immunization to *S. sonnei* through contact with another member of the Enterobacteriaece, *Plesiomonas shigelloides,* with both species sharing a major surface O-antigen that is required for the penetration of human gut epithelial cells[85]. This primarily aquatic organism is commonly found in areas with poor water quality and sanitation and can also cause bouts of diarrhea[86]. It is thought that in areas where *P. shigelloides* is present, in particular the serotype O17, exposure to this organism will impart passive immunity to *S. sonnei*. As countries move towards industrialization with an improvement in sanitation, *P. shigelloides* incidence will decrease[87] along with the local host immunity to both species

carrying the specific O antigen and *S. sonnei* infections will rise[88], reflecting the global trend observed in developing countries.

Recent work looking at the role of the common amoeba *Acanthamoeba* as an environmental host of pathogenic bacteria has proposed another mechanism by which the proportion of *S. sonnei* infections has increased in developing countries. *Acanthamoeba* are very commonly found in water supplies globally, both in industrialized and lower developed countries[89], and have been shown to have the ability to engulf a range of pathogenic bacteria, including *Shigella* species[66,90], protecting the encased organisms from environmental stresses and changes[91]. This mechanism is thought to be analogous to the survival of enteric bacteria in macrophages within the mammalian host[92], with these amoebae additionally acting as a potential site of gene transfer among the encapsulated microbiota[93]. Two papers by Saeed *et al* present evidence that *S. sonnei*, *S. flexneri* and *S. dysenteriae* were actively integrated into *Acanthamoeba castellanii* cysts and survived for up to three weeks[94,95]. The growth of each species was compared, with notable differences between *S. sonnei* and *S. flexneri* when taken up by *A. castellanii. S. sonnei* was readily maintained by the amoeba in temperatures up to 30^{o}C and growth rates were even higher than in the free living bacterial population[95]. On the other hand, at higher temperatures *S. flexneri* reduces the growth and survival of *A. castellanii* and was less likely to be integrated by the amoeba[95], possibly due to the activation of genes involved in cellular invasion by *S. flexneri* that causes cell death[96]. This suggests that *Acanthamoeba* act as an effective source of protection for *S. sonnei*, though not *S. flexneri*, by encapsulating the bacteria and shielding it from antimicrobials and chemical agents[97]. This may be an important factor in the survival of *S. sonnei* populations in developed countries where antibiotic use and sanitation is widespread, and contributed to the emergence of *S. sonnei* in industrializing regions as these practices are increasing.

In addition, as these countries have moved towards improvements in sanitation and the treatment of drinking water there will be an increased use of commercial and industrial disinfectants[98]. A greater tolerance to these chemicals, as well as key differences in the persistence of particular populations or species in these changing environments, may also be contributing to the spread of *S. sonnei*, though this phenomenon has not been previously investigated.

Evidence of explicit differences between *S. sonnei* and *S. flexneri* has been found in two recent studies, highlighting potential environmental and evolutionary differences between species. Work detailing the global evolutionary history of *S. flexneri* found that this species can be classified into seven phylogenetic groups, with distinct geographic and virulence patterns that are not influenced by antimicrobial resistance acquisition[99]. This is in contrast

to *S. sonnei*, which shows a highly clonal evolutionary history influenced by acquisition of resistance mutations[81]. Further differences between *S. sonnei* and *S. flexneri* has been found by looking at the retention and function of the major virulence plasmid, pINV, in these species. The pINV plasmid was found to be less stable in *S. sonnei* than *S. flexneri* due to the absence of a toxin-antitoxin system, GmvAT, which contributes to the retention of this plasmid in *S. flexneri,* particularly at environmental temperatures. The loss of major phenotypic traits, such as virulence and antibiotic resistance, encoded on this plasmid suggests that *S. sonnei* is less adapted to survival outside of a host and will be likely to be transmitted solely by person-to-person contact.

In conclusion, it is clear that an improvement in water quality and an increase in antibiotic use have coincided with the replacement of *S. flexneri* with *S. sonnei* in developing countries. There is evidence for key genetic and epidemiological differences between *Shigella* species that may have contributed to this observed pattern, with *S. sonnei* potentially acquiring a wider range of antimicrobial resistance through the ability to integrate genetic elements. In addition, passive immunity to *S. sonnei* may be decreasing in developing countries as *P. shigelloides* populations declining, and survival of *S. sonnei* may be increased through environmental protection by encapsulation by *A. castellanii*. The combination of these factors appears to be important in increasing the global incidence of *S. sonnei* infections, though there needs to be more work done on genetic and ecological differences between populations of each species to further understand the underlying reasons for this shift.

1.6 Thesis Outline

This thesis comprises projects using both whole genome sequence (WGS) data and *in vitro* experimental approaches to study bacterial genomics and ecology. The main focus of this work is to identify ecological and genetic differences that may have contributed to the recent trend of the replacement of *Shigella flexneri* with *Shigella sonnei* as the primary etiological agent of bacillary dysentery in Vietnam, a pattern that has been observed in many developing countries.

To begin, chapter two describes experimental work examining the survival and growth of *S. sonnei* and *S. flexneri* in cultures containing varying concentrations of calcium hypochlorite and sodium dodecyl sulphate (SDS), two components of chemical disinfectants. The rationale behind these experiments is to determine whether there is any difference in resistance of each species to these chemicals, which may contribute to differential persistence and potential to spread infection in the environment. The increased use of

cleaning products containing these chemicals in developing countries[98] suggests that these biocides could be an important selective pressure on bacterial pathogen populations, and any differences in resistance may play an important role in their spread.

Following on from the work presented in chapter three is a further experimental section focusing on the role of efflux pumps in calcium hypochlorite resistance in *Shigella*, as well as investigating any links to antibiotic resistance phenotypes. There has been support for the importance of efflux pumps in both resistance to antibiotics and chemical antimicrobials, including chlorine[100], and thus it can be postulated that strains resistant to antibiotics may also show a greater resistance to disinfectants.

In chapter four I present work aiming to identify and characterize key chromosomal antibiotic resistance mutations in clinical isolates of *S. sonnei* and *S. flexneri* from Vietnam. Additionally, SNPs that are present in resistant isolates only and those that have evolved independently (homoplasy) will be identified to look for evidence of compensatory mutations. As described in section 1.2, previous studies have shown that the evolution of compensatory mutations can be an important factor in the selection and spread of antibiotic resistance genes. Strains harbouring these resistance genes may have a selective advantage over susceptible bacteria in an environment that contains the specific antibiotic, though these gene modifications can lower the relative fitness, and survival, when the antibiotic is absent. Evidence suggests compensatory mutations can restore fitness levels of resistant bacteria to around that of wild type susceptible strains in antibiotic-free environments, which will increase their selective advantage and potential for propagation. Experimental work from colleagues at the Oxford University Clinical Research Unit (OUCRU) in Ho Chi Minh City, Vietnam has suggested that compensatory mechanisms may be present in Vietnamese populations of *S. sonnei*, particularly with those carrying fluoroquinolone resistance through *gyrA* mutations, and thus I will attempt to identify any putative candidates for compensatory mutations and determine any differences between the *Shigella* species.

Chapter five takes the clinical isolates of Vietnamese *S. sonnei* and *S. flexneri* used in chapter four and with *de novo* assembly of the raw sequence data builds a representative pan genome for each species. This describes the full genomic content of the *Shigella* species in this study, including genes that are shared by all isolates and those found in varying proportions within the population. This method will include local genome content variation that would not be captured through the more commonly used assembly method of aligning raw sequence reads against a reference strain, which may be relatively distant genetically. Of particular interest are genes that are present at high frequencies in one

species but not another that may determine key phenotypic differences between the species and impact the spread of the pathogens, including ecology and virulence.

Finally, chapter six is a general discussion of the previous four original data chapters and how the findings of these projects go towards answering the key aims of the thesis.

In addition, there is a short technical chapter within the supplementary materials describing the inception and purpose of the R function, vcfProcess, for transforming and filtering variant caller output files, such as with GATK or Samtools, in the VCF format, specifically when working on genomes of haploid organisms. The challenge faced when using these variant-calling software on haploid data are two-fold; 1) the software has been predominately designed for diploid data and 2) the output format (.vcf) is not a compatible format to be used as input files for many downstream phylogenetic or population genetic analysis programs. This tool will allow the user to identify and remove heterozygous calls and filter called variant sites based on quality and position in the genome before producing output files in formats that can be readily used with downstream analysis programs.

The work presented here examines ecological, antibiotic resistance and genetic differences between *S. sonnei* and *S. flexneri* that may influence the recent observed pattern of species replacement in developing countries. The aim is to determine how these factors contribute to the increased spread of a particular species or strain, and if the changing environment in these regions may have contributed to the shift in the primary cause of shigellosis infections.

References

1. Wilson, D. J. Insights from Genomics into Bacterial Pathogen Populations. *PLoS Pathog.* **8,** (2012).

2. Cohen, M. L. Changing patterns of infectious disease. *Nature* **406,** 762–767 (2000).

3. Alemayehu, A. Review on Emerging and Re-Emerging Bacterial Zoonotic Diseases. *Am. J. Sci. Res.* **7,** 176–186 (2012).

4. Morens, D. M., Folkers, G. K. & Fauci, A. S. The challenge of emerging and re-emerging infectious diseases. *Nature* **430,** 242–9 (2004).

5. Geard, N. *et al.* The effects of demographic change on disease transmission and vaccine impact in a household structured population. *Epidemics* **13,** 56–64 (2015).

6. Rajagopalan, S. Tuberculosis and Aging: A Global Health Problem. *Clin. Infect. Dis.* **33,** 1034–1039 (2001).

7. Truong, H. M. *et al.* Increases in sexually transmitted infections and sexual risk behaviour without a concurrent increase in HIV incidence among men who have sex with men in San Francisco: a suggestion of HIV serosorting? *Sex. Transm. Infect.* **82,** 461–6 (2006).

8. Gilbart, V. L. *et al.* Factors associated with heterosexual transmission of HIV to individuals without a major risk within England, Wales, and Northern Ireland: a comparison with national probability surveys. *Sex. Transm. Infect.* **82,** 15–20 (2006).

9. Ramjee, G. *et al.* Women and HIV in Sub-Saharan Africa. *AIDS Res. Ther.* **10,** 30 (2013).

10. Eldholm, V. *et al.* Impact of HIV co-infection on the evolution and transmission of multidrug-resistant tuberculosis. *Elife* **5,** 1–19 (2016).

11. Tatem, A. J., Rogers, D. J. & Hay, S. I. Global transport networks and infectious disease spread. *Adv. Parasitol.* **62,** 293–343 (2006).

12. Altekruse, S. F., Cohen, M. L. & Swerdlow, D. L. Emerging Foodborne Diseases. *Emerg. Infect. Dis.* **3,** 285–293 (1997).

13. Jones, B. A. *et al.* Zoonosis emergence linked to agricultural intensification and environmental change. *Proc. Natl. Acad. Sci.* **110,** 8399–8404 (2013).

14. Greger, M. The human/animal interface: emergence and resurgence of zoonotic infectious diseases. *Crit. Rev. Microbiol.* **33,** 243–99 (2007).

15. Altizer, S., Ostfeld, R. S., Johnson, P. T. J., Kutz, S. & Harvell, C. D. Climate Change and Infectious Diseases: From Evidence to a Predictive Framework. *Science (80-.).* **341,** 514–519 (2013).

16. Pascual, M., Bouma, M. J. & Dobson, A. P. Cholera and climate: Revisiting the quantitative evidence. *Microbes Infect.* **4,** 237–245 (2002).

17. Vezzulli, L., Pezzati, E., Brettar, I., Höfle, M. & Pruzzo, C. Effects of Global Warming on Vibrio Ecology. *Microbiol. Spectr.* **3,** (2015).

18. Wu, X., Lu, Y., Zhou, S., Chen, L. & Xu, B. Impact of climate change on human infectious diseases: Empirical evidence and human adaptation. *Environ. Int.* **86,** 14–23 (2016).

19. Gordon, R. J. & Lowy, F. D. Pathogenesis of Methicillin-Resistant *Staphylococcus aureus* Infection. *Clin. Infect. Dis.* **46,** S350–S359 (2008).

20. García, E., Llull, D. & López, R. Functional organization of the gene cluster involved in the synthesis of the pneumococcal capsule. *Int. Microbiol.* **2,** 169–176 (1999).

21. Wilson, J. W. *et al.* Mechanisms of bacterial pathogenicity. *Postgrad. Med. J.* **78,** 216–224 (2002).

22. Finlay, B. B. & Falkow, S. Common themes in microbial pathogenicity revisited. *Microbiol. Mol. Biol. Rev.* **61,** 136–69 (1997).

23. Pupo, G. M., Lan, R. & Reeves, P. R. 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 (2000).

24. Bur, S., Preissner, K. T., Herrmann, M. & Bischoff, M. The Staphylococcus aureus extracellular adherence protein promotes bacterial internalization by keratinocytes independent of fibronectin-binding proteins. *J. Invest. Dermatol.* **133,** 2004–12 (2013).

25. Fields, K. A., Heinzen, R. A. & Carabeo, R. The obligate intracellular lifestyle. *Front. Microbiol.* **2,** 1–2 (2011).

26. Poisot, T., Stanko, M., Miklisová, D. & Morand, S. Facultative and obligate parasite communities exhibit different network properties. *Parasitology* **140,** 1340–5 (2013).

27. Ochman, H. & Moran, N. a. Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis. *Science* **292,** 1096–1099 (2001).

28. Hemminki, E. & Paakkulainen, A. The effect of antibiotics on mortality from infectious diseases in Sweden and Finland. *Am. J. Public Health* **66,** 1180–1184 (1976).

29. Freire-Moran, L. *et al.* Critical shortage of new antibiotics in development against multidrug-resistant bacteria-Time to react is now. *Drug Resist. Updat.* **14,** 118–24 (2011).

30. Martínez, J. L. Antibiotics and antibiotic resistance genes in natural environments. *Science* **321,** 365–7 (2008).

31. MacLean, R. C., Hall, A. R., Perron, G. G. & Buckling, A. The population genetics of antibiotic resistance: integrating molecular mechanisms and treatment contexts. *Nat. Rev. Genet.* **11,** 405–14 (2010).

32. Liu, Y.-Y. *et al.* Emergence of plasmid-mediated colistin resistance mechanism MCR-1 in animals and human beings in China: a microbiological and molecular biological study. *Lancet Infect. Dis.* **16,** 161–168 (2016).

33. Coates, A. R., Halls, G. & Hu, Y. Novel classes of antibiotics or more of the same? *Br. J. Pharmacol.* **163,** 184–194 (2011).

34. Andersson, D. I. The biological cost of mutational antibiotic resistance: any practical

conclusions? *Curr. Opin. Microbiol.* **9,** 461–465 (2006).

35. Andersson, D. I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8,** 260–71 (2010).

36. Maisnier-Patin, S. & Andersson, D. I. Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. *Res. Microbiol.* **155,** 360–9 (2004).

37. Dionisio, F., Conceição, I. C., Marques, A. C. R., Fernandes, L. & Gordo, I. The evolution of a conjugative plasmid and its ability to increase bacterial fitness. *Biol. Lett.* **1,** 250–2 (2005).

38. Vogwill, T. & Maclean, R. C. The genetic basis of the fitness costs of antimicrobial resistance: A meta-analysis approach. *Evol. Appl.* **8,** 284–295 (2015).

39. Furuya, E. Y. & Lowy, F. D. Antimicrobial-resistant bacteria in the community setting. *Nat. Rev. Microbiol.* **4,** 36–45 (2006).

40. Levin, B. R. Models for the spread of resistant pathogens. *Neth. J. Med.* **60,** 58–64 (2002).

41. Zhang, L. & Watson, L. T. Analysis of the fitness effect of compensatory mutations. *HFSP J.* **3,** 47–54 (2009).

42. Schrag, S. J., Perrot, V. & Levin, B. R. Adaptation to the fitness costs of antibiotic resistance in Escherichia coli. *Proc. Biol. Sci.* **264,** 1287–91 (1997).

43. Gagneux, S. *et al.* The Competitive Cost of Antibiotic Resistance in Mycobacterium tuberculosis. **312,** 1944–1947 (2006).

44. Trindade, S. *et al.* Positive epistasis drives the acquisition of multidrug resistance. *PLoS Genet.* **5,** e1000578 (2009).

45. Borrell, S. *et al.* Epistasis between antibiotic resistance mutations drives the evolution of extensively drug-resistant tuberculosis. *Evol. Med. Public Heal.* **2013,** 65–74 (2013).

46. Shiga, K. Ueber den erreger der dysenterie in Japan. *Zentralbl. Bakteriol. Mikrobiol. 23, 599–600* **23,** 23, 599–600 (1898).

47. Peng, J., Yang, J. & Jin, Q. The molecular evolutionary history of Shigella spp. and enteroinvasive Escherichia coli. *Infect. Genet. Evol.* **9,** 147–152 (2009).

48. Faruque, S. M. *et al.* Isolation of Shigella dysenteriae Type 1 and S . flexneri Strains from Surface Waters in Bangladesh : Comparative Molecular Analysis of Environmental Shigella Isolates versus Clinical Strains Isolation of Shigella dysenteriae Type 1 and S . flexneri Strai. *Appl. Environ. Microbiol.* **68,** 3908–3913 (2002).

49. Kotloff, K. L. *et al.* Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull. World Health Organ.* **77,** 651–66 (1999).

50. Njamkepo, E. *et al.* Global phylogeography and evolutionary history of Shigella dysenteriae type 1. *Nat. Microbiol.* 16027 (2016). doi:10.1038/nmicrobiol.2016.27

51. Lan, R. & Reeves, P. R. Escherichia coli in disguise: Molecular origins of Shigella. *Microbes Infect.* **4,** 1125–1132 (2002).

52. DuPont, H. L., Levine, M. M., Hornick, R. B. & Formal, S. B. Inoculum size in shigellosis and implications for expected mode of transmission. *J. Infect. Dis.* **159,** 1126–8 (1989).

53. Ipa, T. Mechanism of Shigella entry into epithelial cells Guy Tran Van Nhieu * and Philippe J Sansonetti. 51–55 (1999).

54. Escobar-Páramo, P., Giudicelli, C., Parsot, C. & Denamur, E. The evolutionary history of Shigella and enteroinvasive Escherichia coli revised. *J. Mol. Evol.* **57,** 140–148 (2003).

55. Sims, G. E. & Kim, S.-H. Whole-genome phylogeny of Escherichia coli/Shigella group by feature frequency profiles (FFPs). *Proc. Natl. Acad. Sci. U. S. A.* **108,** 8329–8334 (2011).

56. Shaikh, N. & Tarr, P. I. Escherichia coli O157 : H7 Shiga Toxin-Encoding Bacteriophages : Integrations , Excisions , Truncations , and Evolutionary Implications Escherichia coli O157 : H7 Shiga Toxin-Encoding Bacteriophages : Integrations , Excisions , Truncations , and Evolution. *J. Bacteriol.* **185,** 3596–3605 (2003).

57. Brenner, D. J., Steigerwalt, A. G., Wathen, H. G., Gross, R. J. & Rowe, B. Confirmation of aerogenic strains of Shigella boydii 13 and further study of Shigella serotypes by DNA relatedness. *J. Clin. Microbiol.* **16,** 432–436 (1982).

58. Brenner, D. O. N. J. *et al.* P oly nucleo tide Sequence Relatedness Among Shigella Species. **23,** 1–7 (2016).

59. Ogura, Y. *et al.* Comparative genomics reveal the mechanism of the parallel evolution of O157 and non-O157 enterohemorrhagic Escherichia coli. *Proc. Natl. Acad. Sci.* **106,** 17939–17944 (2009).

60. van den Beld, M. J. C. & Reubsaet, F. a G. Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli. *Eur. J. Clin. Microbiol. Infect. Dis.* **31,** 899–904 (2012).

61. Reeves, P. R. *et al.* Rates of mutation and host transmission for an escherichia coli clone over 3 years. *PLoS One* **6,** (2011).

62. Zhou, Z. *et al.* Derivation of Escherichia coli O157:H7 from its O55:H7 precursor. *PLoS One* **5,** (2010).

63. Zuo, G., Xu, Z. & Hao, B. Shigella Strains Are Not Clones of Escherichia coli but Sister Species in the Genus Escherichia. *Genomics, Proteomics Bioinforma.* **11,** 61–65 (2013).

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

65. Yang, J. *et al.* Revisiting the Molecular Evolutionary History of Shigella spp. *J. Mol. Evol.* **64,** 71–79 (2007).

66. Thompson, C. N., Duy, P. T. & Baker, S. The Rising Dominance of Shigella sonnei: An Intercontinental Shift in the Etiology of Bacillary Dysentery. *PLoS Negl. Trop. Dis.* **9,** e0003708 (2015).

67. Vinh, H. *et al.* A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation. *BMC Infect. Dis.* **9,** 204 (2009).

68. Dutta, S. *et al.* Shifting serotypes, plasmid profile analysis and antimicrobial resistance pattern of shigellae strains isolated from Kolkata, India during 1995-2000. *Epidemiol. Infect.* **129,** 235–43 (2002).

69. Qu, F. *et al.* Genotypes and antimicrobial profiles of Shigella sonnei isolates from diarrheal patients circulating in Beijing between 2002 and 2007. *Diagn. Microbiol. Infect. Dis.* **74,** 166–70 (2012).

70. Ângela Bernardes Sousa, M. *et al.* Shigella in Brazilian children with acute diarrhoea: Prevalence, antimicrobial resistance and virulence genes. *Mem. Inst. Oswaldo Cruz* **108,** 30–35 (2013).

71. von Seidlein, L. *et al.* A multicentre study of Shigella diarrhoea in six Asian countries: disease burden, clinical manifestations, and microbiology. *PLoS Med.* **3,** e353 (2006).

72. Holt, K. E. *et al.* Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17522–7 (2013).

73. Vinh, H. *et al.* Rapid emergence of third generation cephalosporin resistant Shigella spp. in Southern Vietnam. *J. Med. Microbiol.* **58,** 281–3 (2009).

74. Garrett, V. *et al.* A recurring outbreak of Shigella sonnei among traditionally observant Jewish children in New York City: the risks of daycare and household transmission. *Epidemiol. Infect.* **134,** 1231–6 (2006).

75. Joh, R. I. *et al.* Dynamics of shigellosis epidemics: estimating individual-level transmission and reporting rates from national epidemiologic data sets. *Am. J. Epidemiol.* **178,** 1319–26 (2013).

76. Uren NG, Crake T, L. D. The New England Journal of Medicine as published by New England Journal of Medicine. Downloaded from www.nejm.org on July 28, 2010. For personal use only. No other uses without permission. Copyright © 1994 Massachusetts Medical Society. All rights reserve. (1994).

77. Eu, W. & Th, C. An Outbreak of Shigella sonnei Infection. **1,** 26–29 (1995).

78. Mammina, C., Aleo, A., Romani, C. & Nastasi, A. Shigella sonnei biotype G carrying class 2 integrons in southern Italy: a retrospective typing study by pulsed field gel electrophoresis. *BMC Infect. Dis.* **6,** 117 (2006).

79. García-Fulgueiras, A. *et al.* A large outbreak of Shigella sonnei gastroenteritis associated with consumption of fresh pasteurised milk cheese. *Eur. J. Epidemiol.* **17,** 533–538 (2001).

80. Cohen, D. *et al.* Recent trends in the epidemiology of shigellosis in Israel. *Epidemiol. Infect.* 1–12 (2014). doi:10.1017/S0950268814000260

81. Holt, K. E. *et al.* Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44,** 1056–9 (2012).

82. Ahmed, A. M., Furuta, K., Shimomura, K., Kasama, Y. & Shimamoto, T. Genetic characterization of multidrug resistance in Shigella spp. from Japan. *J. Med. Microbiol.* **55,** 1685–1691 (2006).

83. Toro, C. S. *et al.* Genetic analysis of antibiotic-resistance determinants in multidrug-resistant Shigella strains isolated from Chilean children. *Epidemiol. Infect.* **133,** 81–86 (2005).

84. Ghosh, S., Pazhani, G. P., Niyogi, S. K., Nataro, J. P. & Ramamurthy, T. Genetic characterization of Shigella spp. isolated from diarrhoeal and asymptomatic children. *J. Med. Microbiol.* **63,** 903–910 (2014).

85. Kubler-Kielb, J., Schneerson, R., Mocca, C. & Vinogradov, E. The elucidation of the structure of the core part of the LPS from Plesiomonas shigelloides serotype O17 expressing O-polysaccharide chain identical to the Shigella sonnei O-chain. *Carbohydr. Res.* **343,** 3123–3127 (2008).

86. Tsuchimoto, S., Ohtsubo, H. & Ohtsubo, E. Stable Maintenance of Resistance Plasmid R100. **170,** 1461–1466 (1988).

87. Krovacek, K., Eriksson, L. M., González-Rey, C., Rosinsky, J. & Ciznar, I. Isolation, biochemical and serological characterisation of Plesiomonas shigelloides from freshwater in Northern Europe. *Comp. Immunol. Microbiol. Infect. Dis.* **23,** 45–51 (2000).

88. Sack, D. A., Hoque, A. T., Huq, A. & Etheridge, M. Is protection against shigellosis induced by natural infection with Plesiomonas shigelloides? *Lancet (London, England)* **343,** 1413–1415 (1994).

89. Trabelsi, H. *et al.* Pathogenic free-living amoebae: Epidemiology and clinical review. *Pathol. Biol.* **60,** 399–405 (2012).

90. Jeong, H. J. *et al.* Acanthamoeba: Could it be an environmental host of Shigella? *Exp. Parasitol.* **115,** 181–186 (2007).

91. Aksozek, A., McClellan, K., Howard, K., Niederkorn, J. Y. & Alizadeh, H. Resistance of Acanthamoeba castellanii Cysts to Physical, Chemical, and Radiological Conditions. *J. Parasitol.* **88,** 621–623 (2002).

92. Abd, H., Johansson, T., Golovliov, I. & Sandstro, G. Survival and Growth of Francisella tularensis in Acanthamoeba castellanii. *Society* **69,** 600–606 (2003).

93. Goebel, W. & Gross, R. Intracellular survival strategies of mutualistic and parasitic prokaryotes. *Trends Microbiol.* **9,** 267–273 (2001).

94. Saeed, A., Abd, H., Edvinsson, B. & Sandström, G. Acanthamoeba castellanii an environmental host for Shigella dysenteriae and Shigella sonnei. *Arch. Microbiol.* **191,** 83–88 (2008).

95. Saeed, A., Johansson, D., Sandström, G. & Abd, H. Temperature Depended Role of Shigella flexneri Invasion Plasmid on the Interaction with Acanthamoeba castellanii. *Int. J. Microbiol.* **2012,** 917031 (2012).

96. Zychlinsky, A. *et al.* In vivo apoptosis in Shigella flexneri infections. *Infect. Immun.* **64,** 5357–5365 (1996).

97. King, C. H., Shotts, E. B., Wooley, R. E. & Porter, K. G. Survival of coliforms and bacterial pathogens within protozoa during chlorination. *Appl. Environ. Microbiol.* **54,** 3023–3033 (1988).

98. Trasande, L. *et al.* How developing nations can protect children from hazardous chemical exposures while sustaining economic growth. *Health Aff.* **30,** 2400–2409 (2011).

99. Connor, T. R. *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri. *Elife* **4,** 1–16 (2015).

100. Yuan, Q. Bin, Guo, M. T. & Yang, J. Fate of antibiotic resistant bacteria and genes during wastewater chlorination: Implication for antibiotic resistance control. *PLoS One* **10,** 1–11 (2015).

Chapter 2

Chlorine and detergent resistance *Shigella sonnei* and *Shigella flexneri*

2.1 Abstract

The disinfection of microbial pathogens with industrial and commercial cleaning products is often the first line of defence against the spread of disease. *Shigella* is an important human pathogen that is spread through faecal-oral transmission, with evidence of outbreaks from environmental sources. Recently in developing countries, there has been a shift in the dominant species of *Shigella* isolated in patients, from *S. flexneri* to *S. sonnei*. Differential sensitivity to disinfectants between these species has been proposed to contribute to this species replacement.

In this chapter, I present *in vitro* experimental work examining resistance Vietnamese *S. sonnei* and *S. flexneri* to two common components of antibacterial products, calcium hypochlorite (producing free chlorine) and the detergent, sodium dodecyl sulphate (SDS), through individual inactivation and pairwise competition assays. Results indicated that Vietnamese *S. sonnei* was able to survive in higher concentrations of calcium hypochlorite than Vietnamese *S. flexneri,* and these differences were likely due to local adaptation of the microbial populations as these differences were not observed in European isolates of each species. Competitive growth assays between the Vietnamese strains also revealed evidence of an antagonistic action on the growth of *S. flexneri* by *S. sonnei* in the presence of calcium hypochlorite. SDS appeared to have little to no direct antimicrobial effect on bacterial growth.

The results of these experiments suggest that Vietnamese *S. sonnei* are able to survive higher concentrations of chlorine-based disinfection than Vietnamese *S. flexneri*. With the increasing sanitation and use of cleaning products in developing countries, these differences in sensitivity to disinfectants may be contributing to the replacement of *S. flexneri* with *S. sonnei* in these regions.

2.2 Introduction

Shigella is a genus of Gram-negative pathogenic bacteria causing bacillary dysentery (shigellosis) in humans, reportedly accounting for over 165 million infections that result in around 1 million deaths each year[1]. Recently there has been evidence of a shift in the primary etiological agent of shigellosis in many developing countries, with infections caused by *Shigella sonnei* rapidly replacing *S. flexneri* in economically developing countries such as Malaysia, China and Vietnam over the last 30 years[2–4]. Understanding the genetic and phenotypic factors that contribute to the spread of these pathogens is vital for preventing infection and implementing effective management strategies.

While *Shigella* is predominately spread through faecal-oral transmission[5], there is evidence that the bacteria can survive for a significant time in contaminated water and food[6–9], as well as being deposited on fomites[10]. There have been examples of *Shigella* outbreaks that have been traced back to external sources, including contaminated tomatoes[11], hand washing facilities in a microbiology laboratory[12], and through recreational swimming in a lake[13]. Most research on infectious diseases is focused on treatment, such as the efficacy of antibiotics However, understanding the epidemiology of transmission of *Shigella* may inform effective interventions to prevent the spread of the pathogen before infection.

Through survival on a host and external deposition in the environment, the bacteria will be in contact with disinfectants frequently used as antimicrobial agents. Chemical chlorination is a commonly used treatment for disinfecting reclaimed wastewater and drinking water supplies, as well as swimming pools[14,15]. At high concentrations chlorine will almost certainly kill most microbes, though due to the potential toxic nature of the compounds to humans they are often used in lower concentrations when treating drinking water and food. At these levels there is the potential for the bacteria to survive and develop resistance to these compounds[16]. Owing to a number of documented cases of bacterial disease transmission from sources that would likely have been treated with chlorine compounds, studies have attempted to quantify bacterial survival and growth in the presence of chlorine. There is evidence that chlorine disinfectant resistance is affected by attachment to surfaces, biofilm formation, nutrient availability and type of chlorine compound[17–19]. There is also variation in chlorine resistance between bacterial species, with Gram-negative microbes showing a greater sensitivity to chlorine treatments than Gram-positive species[16]. The availability of organic matter can also effect the antimicrobial efficacy

of chlorine as the levels of free chlorine in a solution will reduce with oxidation of these organic compounds[16,20].

Although there is an abundant use of chlorine compounds in disinfectant products, their antimicrobial activity is not yet fully understood, with few recent studies attempting to describe potential mechanisms. Chlorine is added to water supplies for disinfectant purposes in the form of soluble compounds (such as calcium hypochlorite, $Ca(ClO)_2$) or gaseous chlorine (ClO_2). In an aqueous solution there are two forms of chlorine, referred to collectively as free chlorine, un-ionised hypochlorous acid (HClO) and hypochlorite ions (ClO^-)[21]. Calcium hypochlorite is the most commonly used compound for chlorination of water as it is very stable and easily dissolved, and there is evidence of antimicrobial properties when in solution[22,23]. The reaction of calcium hypochlorite and water is:

(1) $Ca(ClO)_2 + 2H_2O \longrightarrow 2HClO + Ca^{2+} + + 2OH^-$

calcium hypochlorite + water —> hypochlorous acid + calcium + hydroxide

It has been proposed that the mechanism by which chlorine kills microorganisms is through interactions with the cellular envelope, namely the disruption of the cell wall and damage to components of the cellular membrane[16,24]. The particular role of the cell wall in chlorine inactivation is not yet fully resolved, though it has been suggested that due to the strong oxidative properties of chlorine, when bound to the bacteria, free chlorine will target specific components of the cell wall and compromise the integrity[25]. Subsequent interaction with the cell membrane will cause changes to the permeability and the release of important constituents from within the cell in both gram-negative and gram-positive bacteria, though there is less reported membrane damage to Gram-positive species, likely due to chlorine having a greater effect on the outer membrane, present only in Gram-negative bacteria[16,26]. This change in permeability can alter transport pathways across the membrane and affect metabolic processes within the cell[21], such as ATP production, which can ultimately lead to cell death[27]. It has also been shown that chlorine compounds such, as hypochlorites, can directly react with nucleotides[28,29].

Sodium dodecyl sulphate (SDS), an anionic surfactant, is one of the most commonly used agents in a variety of household and industrial detergents, as well as personal hygiene products[30]. It is primarily used due to its foaming ability as well as its biodegradability by a number of soil and aquatic bacteria[31,32]. Direct antimicrobial activity by SDS is uncertain, with the surfactant often combined with a toxic or

antibiotic compound in cleaning products such as a cationic detergent. However, there is evidence that SDS can have some direct effect on biofilm formation[33] and, importantly, induce shock response proteins in enteric pathogens[34,35]. As the human intestinal tract is high in bile salts, which can act as an antimicrobial much the same as chemical detergents[36], resistance to SDS can also effect the sensitivity to these compounds through similar response mechanisms[35]. The ubiquity of SDS use in cleaning products and the evidence for accumulation in the environment[30] suggests that the ability to persist and grow in high concentrations of SDS can influence the survival and transmission of pathogens .

Though there has been previous work looking at the action of chlorine and detergents on *Shigella* and related enteric bacteria, to date there have been no studies looking specifically at differences in chlorine and SDS tolerance between species of *Shigella*, and particularly at differences in sub-lethal doses of these chemical disinfectants. Differential persistence and growth in the presence of these common cleaning agents may influence the effectiveness of these antimicrobial treatments, and thus if a species is able to survive for longer or against greater concentrations in the environment, the chance of opportunistic infection by that pathogen might increase. Through improvements in water quality and sanitation in many developing countries[37], there is an increased use of chemical antimicrobials and these can exert selective pressure on pathogenic bacterial populations, with evidence of disinfectant use influencing both species survival[19,38], virulence[17] and antibiotic resistance acquisition[14].

The primary objective of the experimental work presented here is to examine the resistance of *Shigella sonnei* and *Shigella flexneri* to chlorine (in the form of calcium hypochlorite) and SDS by measuring bacterial survival against varying concentrations of these disinfectants. In particular, the experiments aim to reveal whether the tested biocides will have a significantly different efficacy against Vietnamese populations of these pathogens, which has been hypothesised to contribute to the observed species replacement in this region previously described[39]. Three clinical isolates of *S. sonnei* and *S. flexneri* from Vietnamese collections were tested for resistance to disinfectants through calcium hypochlorite and SDS inactivation assays. Additionally, these experiments were conducted with three historical strains of both *S. flexneri* and *S. sonnei* from Europe and three strains of European enteropathogenic *E. coli* (EPEC) to determine whether any potential differential survival between Vietnamese strains of *S. sonnei* and *S. flexneri* was due

to local, recent adaptation in these Asian populations, with the European *Shigella* and EPEC strains representing proxies for the likely ancestral phenotypes. Finally, the three Vietnamese strains of *S. sonnei* and *S. flexneri* were grown together in all-against-all co-cultures with solutions containing varying sub-lethal concentrations of calcium hypochlorite and SDS to determine the competitive fitness of each species under these conditions.

2.3 Materials and methods

Bacterial strains

Clinical strains of *S. sonnei, S. flexneri* and EPEC *E. coli* taken from laboratory collections were used in this study. Vietnamese *Shigella* strains *(S. sonnei* DE1208, MS004 and EG430, and *S. flexneri* DE0350, MS0052 and EG419) were obtained from the Oxford University Clinical Research Unit (OUCRU) in Ho Chi Minh City, Vietnam. European samples (*S. sonnei* strains 25410, 88-83, 43-74; *S. flexneri* strains 9-63, 12-66, 262-78; and enteropathogenic *E. coli* (EPEC) strains 12823, 13866 and 27592) were acquired from the group of Professor Francois-Xavier Weill in the Institute Pasteur, Paris. Further details of each strain is given in **Table 1.** Vietnamese isolates were selected as reprehensive strains within the region, with all *S. sonnei* samples part of the global lineage III and *S. flexneri* belonging to the serotype 2a type. European *Shigella* and EPEC strains were chosen where WGS data and phenotypic information was also available for further genomic analysis.

Inocula preparation

Stock cultures of bacteria strains were maintained in frozen liquid LB broth supplemented with 10% (v/v) glycerol at -80ºC. To prepare inocula cultures for trials, a small amount of the frozen bacterial stock was transferred via sterile loop to solid LB agar plates and incubated at 37ºC overnight. A single colony was then picked and transferred to 20ml of M9 minimum media (Sigma-Aldrich, Dorset UK) supplemented with 12.5mg/l nicotinic acid and 20mg/l tryptophan[40] and grown overnight in a shaking incubator (37ºC, 250 rpm). Cultures were then diluted with M9 media to around 0.5 OD_{600nm} (around 6-7 x 10^6 cells/ml) and grown for a further two hours to achieve log phase growth of bacteria. Initial cell concentrations were recorded through bacterial population counts and cultures used to inoculate test media.

Species	Strain	Abbreviation	Origin	Other strain information	Available antibiotic resistance phenotype
Shigella flexneri	12-66	S.f12-66	Europe	Serotype 2a	Pan-susceptible
Shigella flexneri	262-78	S.f262-78	Europe	Serotype 1a	Pan-susceptible
Shigella flexneri	9-63	S.f9-63	Europe	Serotype 2a	Streptomycin, Spectinomycin, Sulfonamides
Shigella flexneri	MS0052	S.f-MS	Vietnam	Serotype 2a	Pan-susceptible
Shigella flexneri	DE0350	S.f-DE	Vietnam	Serotype 2a	Fluoroquinolones
Shigella flexneri	EG419	S.f-EG	Vietnam	Serotype 2a	Fluoroquinolones
Shigella sonnei	54210	S.s54210	Europe	Lineage III	Pan-susceptible
Shigella sonnei	88-83	S.s88-83	Europe	Lineage III	Pan-susceptible
Shigella sonnei	43-74	S.s43-74	Europe	Lineage I	Pan-susceptible
Shigella sonnei	MS004	S.sMS	Vietnam	Lineage III	Pan-susceptible
Shigella sonnei	DE1208	S.sDE	Vietnam	Lineage III	Pan-susceptible
Shigella sonnei	EG0430	S. sEG	Vietnam	Lineage III	Fluoroquinolones, Cephalosporins
Escherichia coli	12823	E.c12	Europe	EPEC	Amoxicillin, Sulfonamides, Chloramphenicol, Tetracycline
Escherichia coli	13866	E.c13	Europe	EPEC	Amoxicillin, Sulfonamides
Escherichia coli	27592	E.c27	Europe	EPEC	Amoxicillin, Sulfonamides

Table 1. Strains used in this study. Antibiotic susceptibility phenotypes were determined by available metadata sent with the strains from MIC experiments.

Bacterial population counts

Initial cell concentration and bacterial survival within each tested condition was determined through plate counts of colony forming units (CFU/ml) in serial dilutions from 10^0 to 10^{-5} at specified time points. 10µl triplicates for each dilution were plated onto solid LB agar plates (supplemented with 0.125% Congo red (Sigma-Aldrich, Dorset UK) for virulence plasmid retention[41]) and incubated for 24-48 hours at 37°C. The resulting colonies were counted by eye at appropriate dilutions, with the mean of the triplicate readings multiplied to the correct power to estimate the average colony count in each tested condition.

Test media preparation

Bacterial survival was tested in the presence of varying concentrations of calcium hypochlorite and sodium dodecyl sulphate (SDS). Tests with calcium hypochlorite were conducted by adding a measured weight of pure calcium hypochlorite ($Ca(ClO)_2$) (Sigma-Aldrich, Dorset UK) to a set volume of liquid M9 minimum media to make a total of 8ml of solution before inoculation with the bacteria. The pH of the M9 media had been measured and adjusted to pH 6 by adding small amounts of dilute hydrochloric acid (HCl) as the calcium hypochlorite will raise the pH of the solution when added. Calcium hypochlorite was chosen as it has higher available chlorine content when in solution and is the most commonly used compound for water chlorination[42]. SDS (>99% reagent, Sigma-Aldrich, Dorset UK) was added as an aqueous solution in appropriate volumes and solutions made up to the final pre-inocula test volume of 8ml with M9 minimum media at a pH of 7. All liquid reagents and solutions were maintained at a temperature of 37°C.

Inactivation experiments

Resistance to calcium hypochlorite and SDS was investigated by carrying out inactivation assays on all Vietnamese and European *Shigella* and European *E. coli* strains. All strains were tested in triplicate. For each separate assay, 8ml of test media was prepared in 15ml centrifuge tubes as described above to concentrations of 0, 2.5, 5, 10 and 20 mg/l of calcium hypochlorite and 0, 15, 22.5 and 30 % SDS in M9 minimum media. Each test media was then inoculated with 2ml of the bacterial strain inoculum to achieve an initial total cell count of ~10^{6-7} CFU/ml. Each tube was incubated whilst shaking vigorously (37°C, 250rpm) for 5, 15 and 30 minutes. Calcium hypochlorite solutions were neutralised by adding 100µl of 0.1M sodium

thiosulphate ($Na_2S_2O_3$) (Sigma-Aldrich, Dorset UK) to each chlorinated solution to bind to and neutralise the available free chlorine. SDS solutions were neutralised by washing, with three cycles of centrifugation and resuspension in PBS buffer solution. At each recorded time point the experiment tubes were transferred from incubation to the centrifuge and spun on a medium speed (5000 RPM or 1677 x *g*) for 3 mins. The supernatant was discarded and the pellet re-suspended in 10ml of phosphate-buffered saline (PBS; 1 mM KH_2PO_4, 0.27 mM KCl, 13.7 mM NaCl, pH 7.4; Sigma-Aldrich, Dorset UK) at room temperature. The centrifugation and PBS washing stage was repeated and the final pellet re-suspended in 10 ml M9 minimum media before plating for bacterial population counts.

In vitro *competition assays*

Competitive growth experiments were conducted with each of the three Vietnamese strains of *S. sonnei* against each Vietnamese strain of *S. flexneri* in the presence of sub-lethal concentrations of calcium hypochlorite (0, 1.5, 2.5, and 5 mg/l $Ca(OCl)_2$) and SDS detergent (0, 15, and 22.5 % SDS). As with the inactivation experiments, 8ml of each test media was prepared in 15ml centrifuge tubes with 1ml of *S. sonnei* and 1ml of *S. flexneri* inocula added to give a total concentration of ~10^{6-7} CFU/ml of bacteria. The media were then placed in a shaking incubator (37^{o}C, 250rpm) for 24 hours before neutralization of the free chlorine or SDS using the same methods described in the inactivation experiments. Survival counts were estimated in CFU/ml with the bacterial population count technique. To differentiate between colonies of either species, colonies were grown on LB plates supplemented with 20mg/ml X-gal (Sigma-Aldrich, Dorset UK) and 0.1M IPTG (Bayer AG, Wuppertal, Germany) to perform blue/white screening. *S. sonnei* strains possess the ability to ferment lactose (slowly) through an active lacZ operon whereas *S. flexneri* do not possess a copy of the *lacZ* gene[43]. Plate counts were conducted after 48 hours to allow for the slow fermentation of lactose by *S. sonnei.*

Statistical analysis

Average CFU/ml counts obtained from experiments on individual strains were combined for each species and location group to give the mean survival counts. These values were then used to calculate the survival fraction of the starting population count to plot graphs using Microsoft Excel of the fraction of surviving bacteria after disinfection given a particular concentration or contact time, and to model inactivation kinetics and conduct further statistical analyses in *R*[44]*.* Where

there was no observable bacterial growth on plates, the population count was set to the highest undetectable value (N = 10 CFU/ml) for statistical modelling.

Inactivation kinetics

Chlorine disinfection of bacteria is often considered to follow a pseudo-first-order kinetic model of inactivation, Chick's Law, which describes the survival of microorganisms after treatment as a function of contact time[45].

$$ln\frac{N}{N_o} = -kt$$

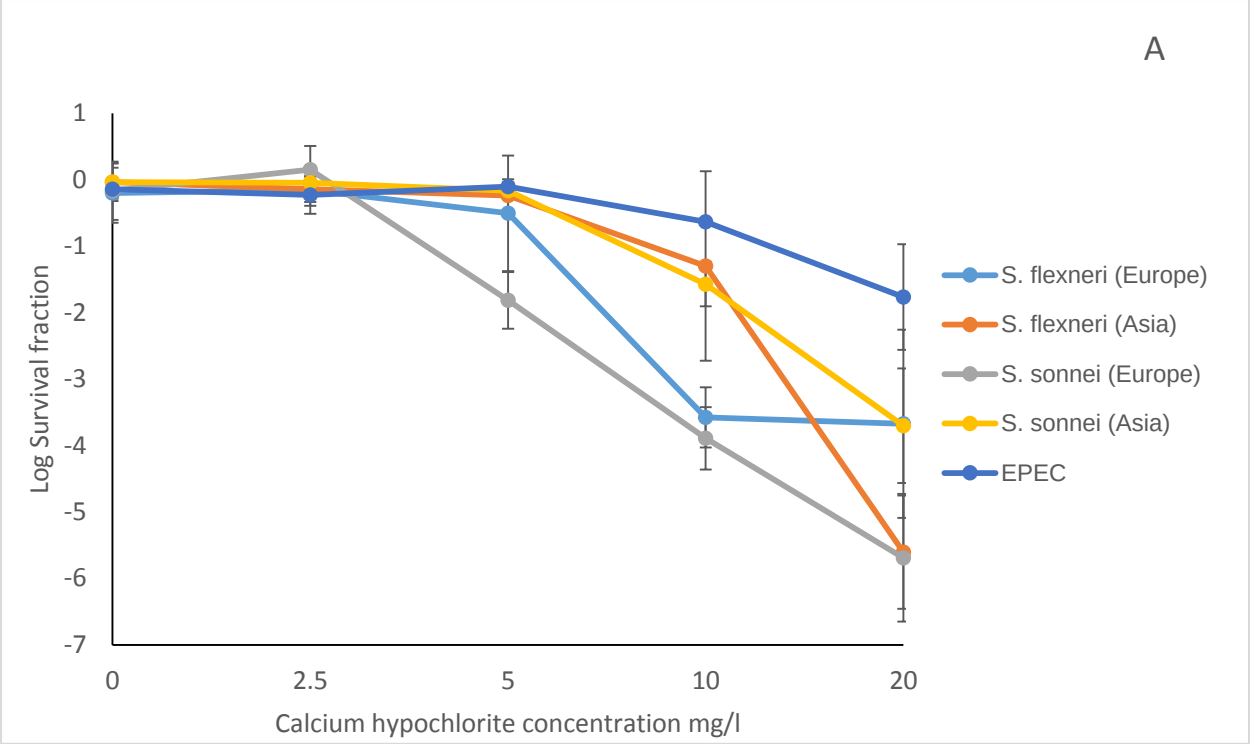
N = Number of bacteria at time *t*
N_o = Initial number of bacteria
t = Contact time
k = Rate constant of disinfection

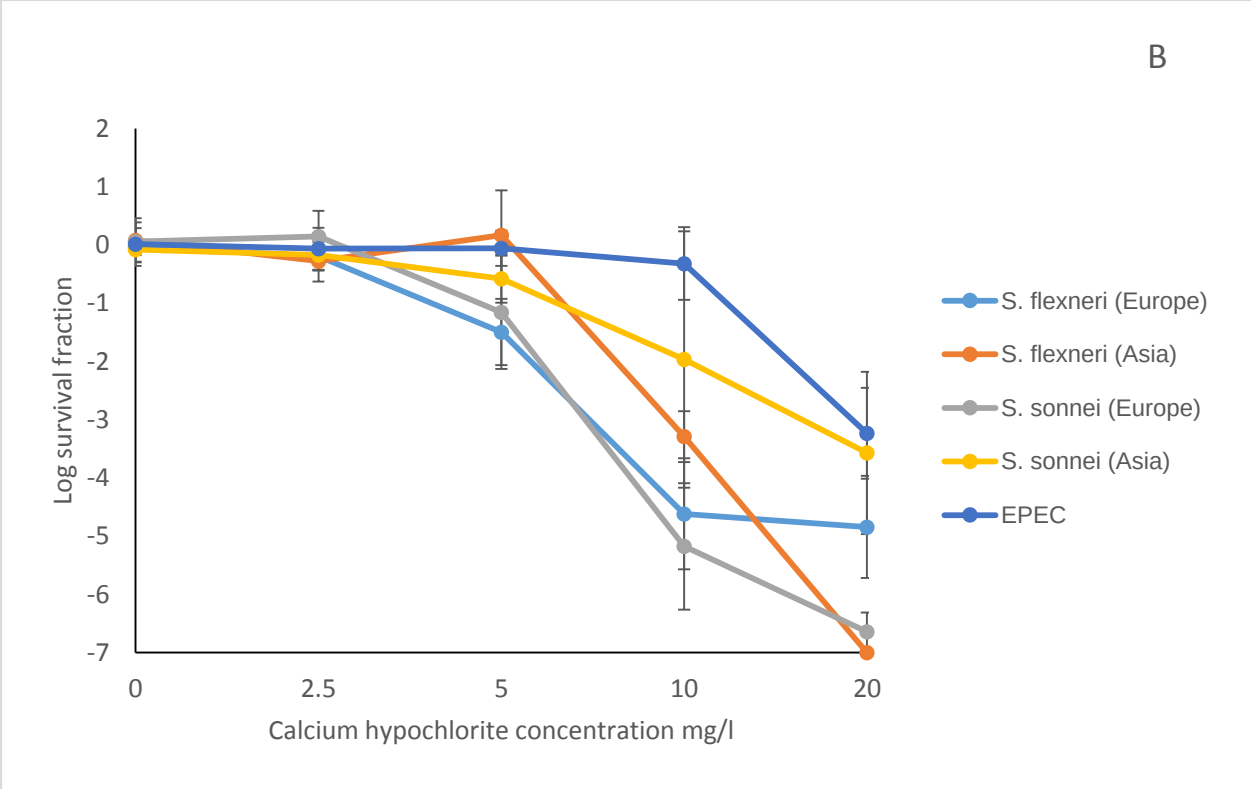
Chick's law of inactivation was later expanded by H.E. Watson to relate *k* to the concentration of disinfectant (*C*) at a dilution ^{n} (almost always 1). This model has the assumptions that there is a constant rate of chlorination, temperature and pH over time, and that the mode of disinfection is single-hit and single-site. This is known as the Chick-Watson model[46].

$$ln\frac{N}{N_o} = -kC^n t$$

N = Number of bacteria at time *t*
N_o = Initial number of bacteria
t = Contact time
C^n = Concentration of disinfection at dilution n
k = Chick-Watson coefficient of specific lethality (rate constant of disinfection)

CT values (C^nt) are often used as a comparison of the chlorine dosage required for inactivation of a given order of magnitude (i.e. log^{-2}, log^{-3}, ...) of microorganisms ($CT_{99\%}$, $CT_{99.9\%,}$ …) as a product of the contact time and concentration. In these experiments, a constant rate of calcium hypochlorite concentration was assumed for up to 30 minutes of contact time for simplification of the analysis, with C^n remaining constant for each calcium hypochlorite concentration tested. It has been reported


A
1
0
-1
-2
-3
-4
-5
-6
-7
Log Survival fraction
0
2.5
5
10
20
Calcium hypochlorite concentration mg/l
S. flexneri (Europe)
S. flexneri (Asia)
S. sonnei (Europe)
S. sonnei (Asia)
EPEC


B
2
1
0
-1
-2
-3
-4
-5
-6
-7
Log survival fraction
0
2.5
5
10
20
Calcium hypochlorite concentration mg/l
S. flexneri (Europe)
S. flexneri (Asia)
S. sonnei (Europe)
S. sonnei (Asia)
EPEC

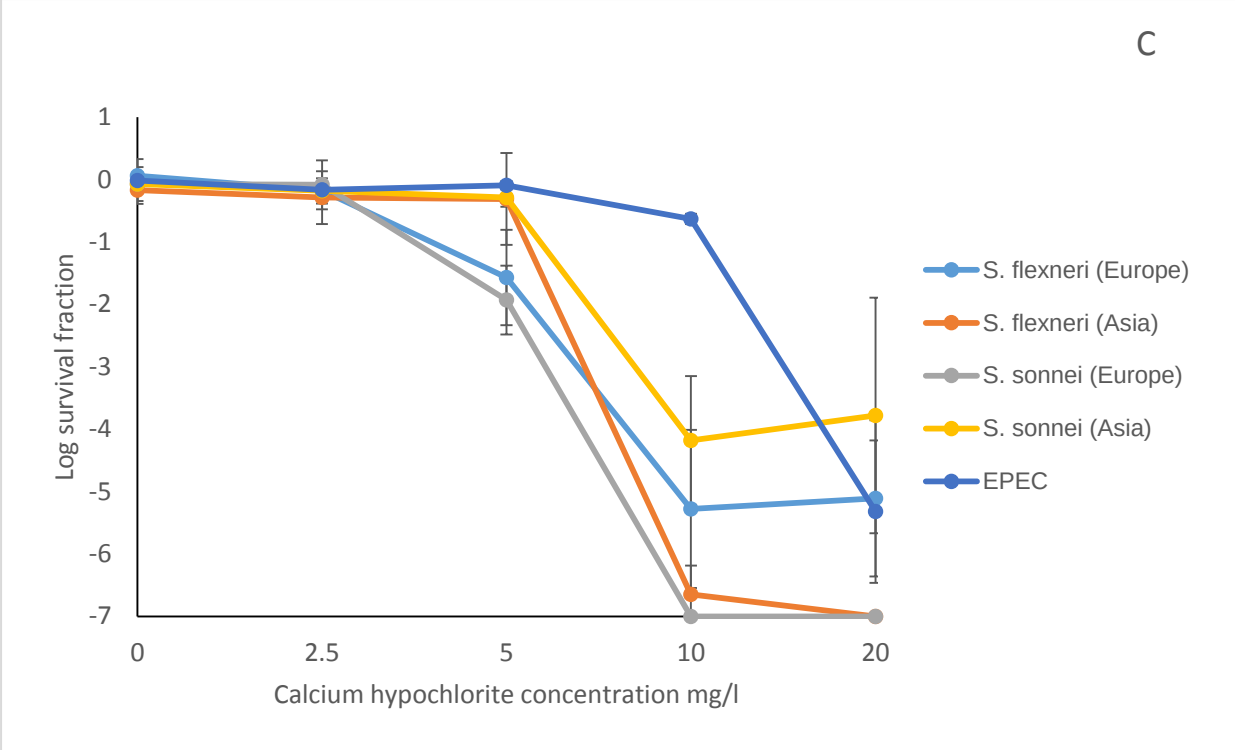


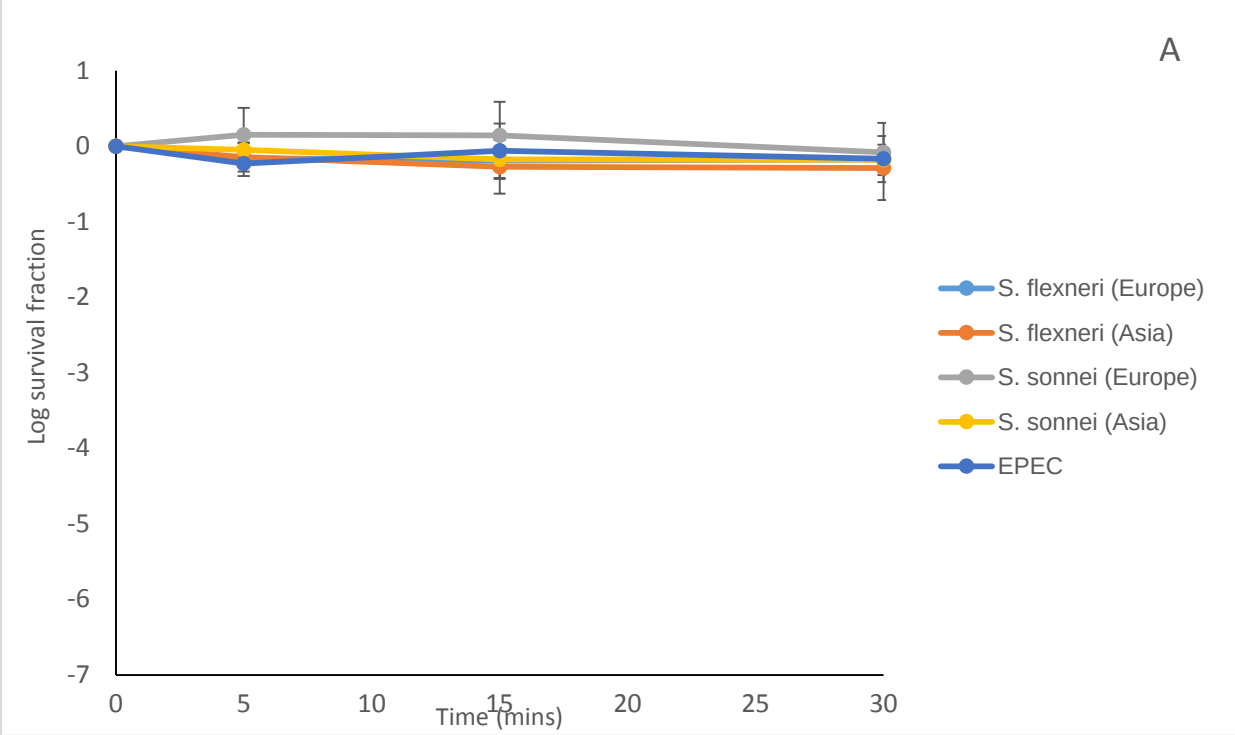
Figure 1. Survival of *Shigella* species and *EPEC* against concentration of calcium hypochlorite. The log survival fraction (N/N_o) is shown at each recorded time point (A) 5 minutes, (B) 15 minutes, and (C) 30 minutes. Error bars denote the standard error at each recorded point.

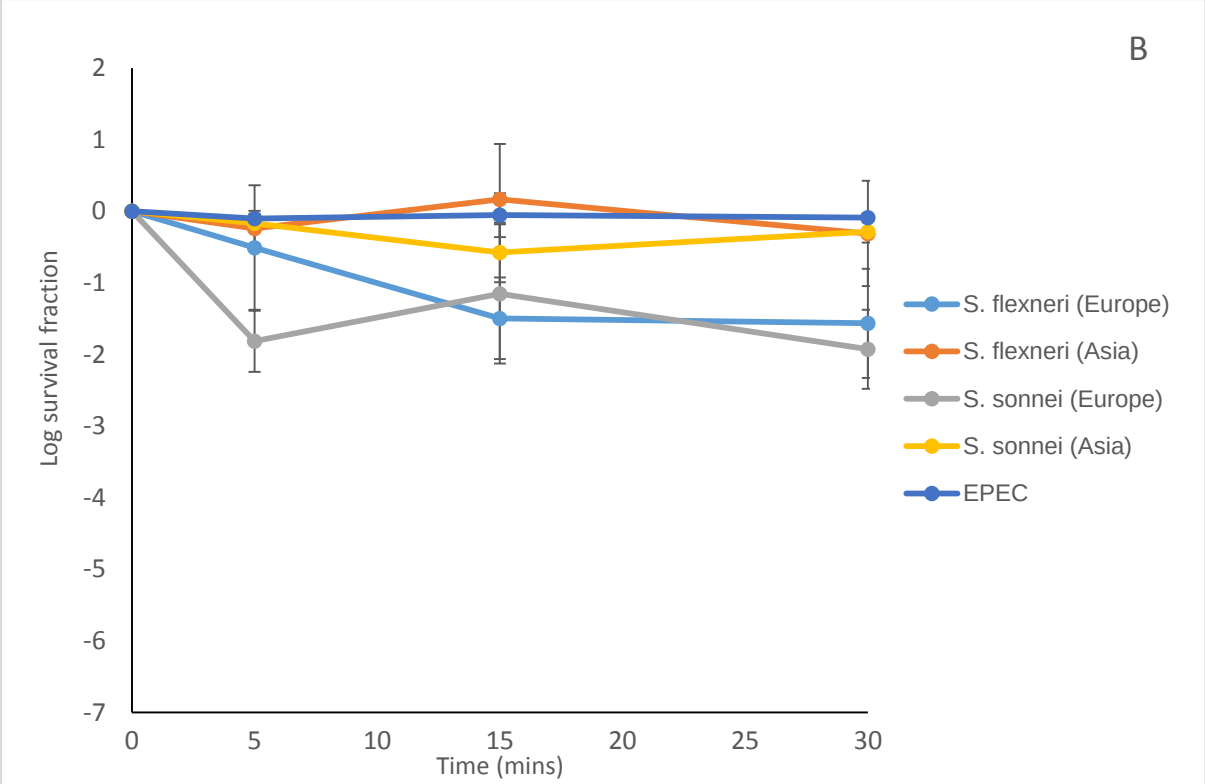
that, although the amount of free chlorine in as solution will reduce in this time, it will do so linearly[47], thus, whilst there may be an overestimation of the absolute CT values in this study, comparison of these values between species groups is valid.

2.4 Results

Inactivation by calcium hypochlorite

To investigate the effect of chlorine on the survival of Vietnamese and European *S. sonnei* and *S. flexneri,* and European enteropathogenic *E. coli* (EPEC), inactivation experiments were carried out by adding live, log-phase cultures of each species to calcium hypochlorite solutions ranging in concentrations from 2.5 to 20 mg/l, with bacterial cell colony counts taken at 5, 15 and 30 minutes. Experiments were stopped at 30 minutes as, at this time, the levels of free chlorine in the system from the added calcium hypochlorite has been shown not to reduce significantly. The log


A
1
0
-1
-2
-3
-4
-5
-6
-7
Log survival fraction
0
5
10
15
20
25
30
Time (mins)
S. flexneri (Europe)
S. flexneri (Asia)
S. sonnei (Europe)
S. sonnei (Asia)
EPEC


B
2
1
0
-1
-2
-3
-4
-5
-6
-7
Log survival fraction
0
5
10
15
20
25
30
Time (mins)
S. flexneri (Europe)
S. flexneri (Asia)
S. sonnei (Europe)
S. sonnei (Asia)
EPEC

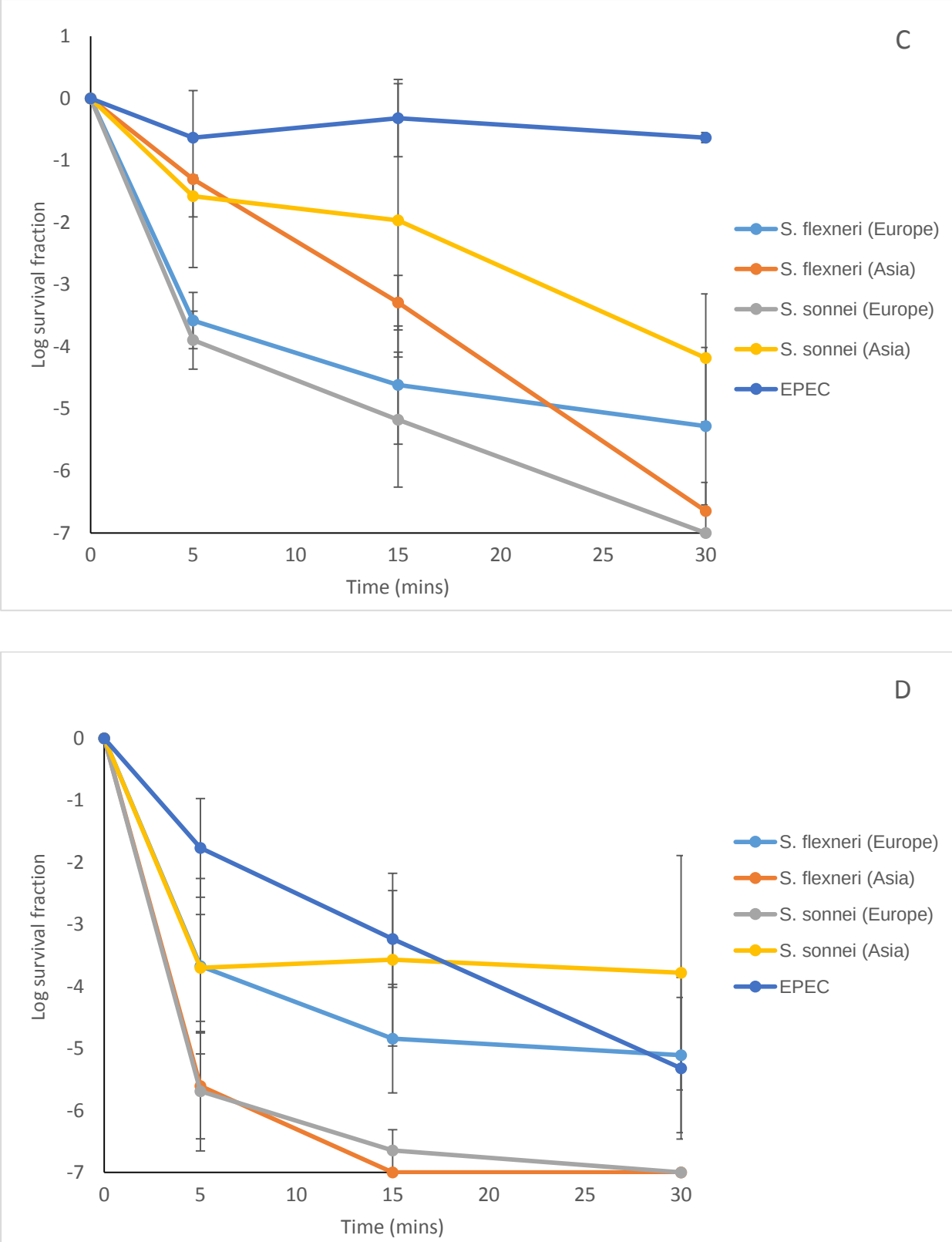


Figure 2. Survival of *Shigella* species and *EPEC* against contact time in minutes in calcium hypochlorite. The log survival fraction (N/N_o) is shown for each tested concentration of calcium hypochlorite (A) 2.5 mg/l, (B) 5 mg/l, (C) 10 mg/l, and (D) 20 mg/l. Error bars denote the standard error at each recorded point.

survival fraction of each species group as a product of calcium hypochlorite concentration and of time is shown in **Figure 1** and **Figure 2** respectively.

For all tested species groups an increase in calcium hypochlorite concentration decreased the survival fraction of bacteria, and in most cases there was a reduction in the number of surviving bacteria as time increased, particularly at higher concentrations of disinfectant (**Figure 1)**. At lower concentrations (2.5 – 5 mg/l) though, there was only a decrease in bacterial survival (up to 99% killed) recorded in European *S. sonnei* and *S. flexneri* (**Figure 2B)** within the contact time of the experiments. EPEC strains were able to tolerate higher calcium hypochlorite concentrations (10 mg/l) with Vietnamese *S. sonnei* exhibiting the highest resistance of the *Shigella* groups (**Figure 2C)**. This was also true for the highest concentration of calcium hypochlorite (20 mg/l) (**Figure 2D)**, though the populations of all species groups were reduced by over 99.9% after 30 minutes of contact time. At these higher concentrations of calcium hypochlorite, European *S. sonnei* and Vietnamese *S. flexneri* were the most sensitive to disinfection, with near or completely undetectable CFU counts. For all groups there was greater within group variance of the survival fraction across all time points at higher concentrations (10-20mg/l), This could be due to some variation in strain tolerance within each species group at these concentrations, along with stochasticity in bacterial response to chlorine disinfection. It is not likely due to innate fitness differences between strains as this level of variation is not seen at low concentrations of chlorine (**Figure 2A)**, or in controls (**Supplementary materials).**

Chlorine inactivation kinetics

To compare the sensitivity of each species group to chlorination, the Chick-Watson disinfection model was employed to determine the rate of disinfection and the $\log^{-3}$ inactivation $CT_{99.9\%}$ value. In these experiments the rate of disinfection follows a first-order linear kinetic rate of decay **(Figure 3)**, and thus the rate constant *k* can be found by finding the linear least-squares regression of the natural log of the survival fraction ($ln\ N/N_o$) plotted against the CT value ($C^n t$), where *a* is the *y* intercept and b is the slope constant:

$$ln\frac{N}{N_o} = -kC^n t$$

$$k = y = a + bx$$

This can then be used to find the CT value for the 99.9% (3-log) inactivation of each species group, where the 99.9% reduction in bacterial colony counts from the initial number will be the natural log of the survival fraction 0.001:

$$ln\frac{N}{N_o} = ln(0.001)$$

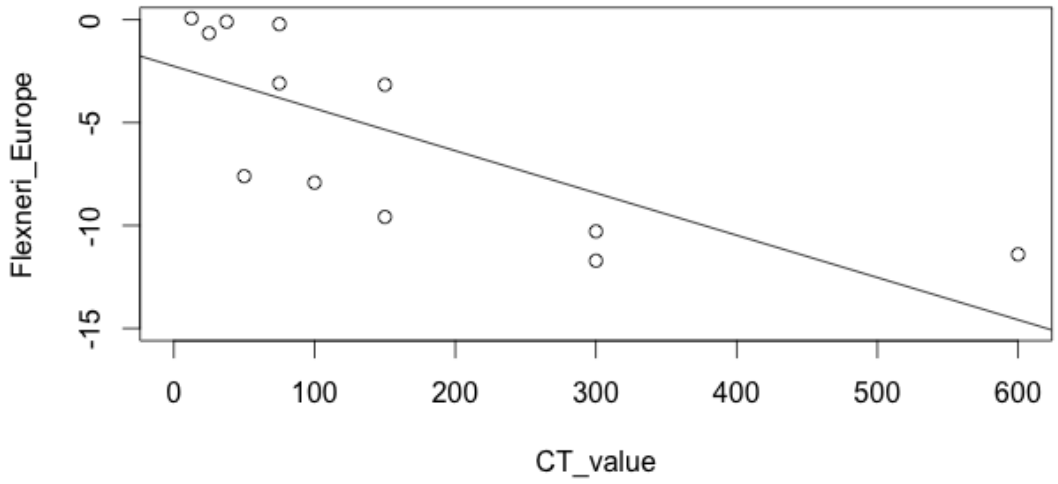
$$ln\ (0.001) = -kC^n t_{99.9\%}$$

Given $k = ln\ (0.001)$ this can be rearranged to find x ($C^n t$):

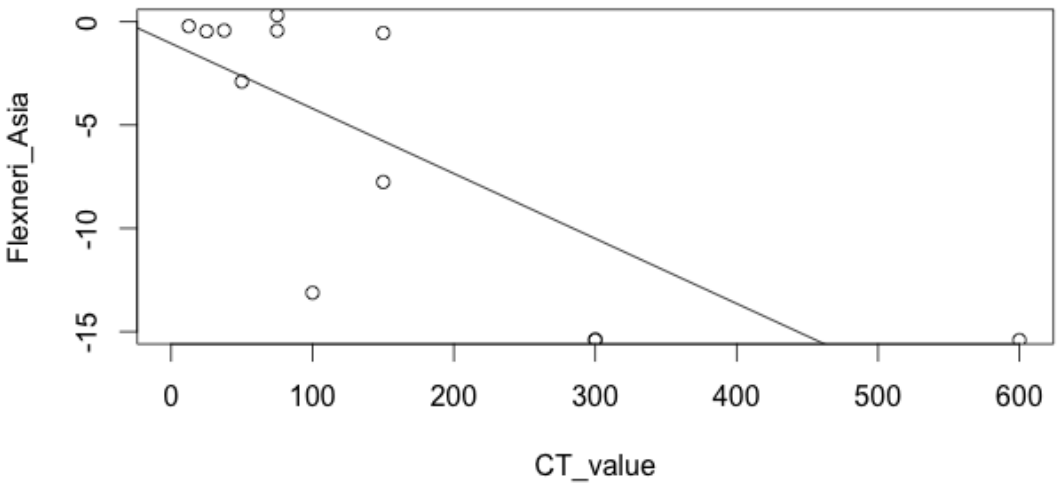
$$C^n t_{99.9\%} = \frac{(ln\ (0.001) - \ a)}{b}$$

Plots of the natural log survival fraction, $ln(N/N_0)$, against $C^n t$ for each species group with the modelled rate constant are illustrated in **Figures 3A-E**. All calcium hypochlorite concentration data were combined for each species group in the analysis as the use of $C^n t$ will explain the survival fraction as a product of the dosage and so allows for differences in experimental concentrations or contact time. The equation for the rate constant of disinfection k and $CT_{99.9\%}$ values for each species group are given in **Table 2**. Linear regression models for each species group indicated a goodness of fit, with adjusted R^2 ranging from 50.65 – 73.45% and all p values ≤0.05.

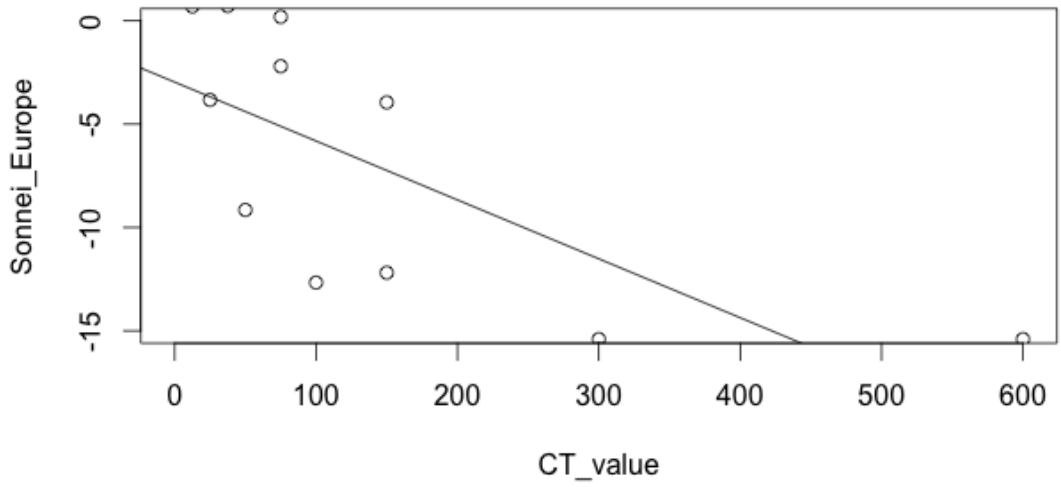
(A)


Flexneri_Europe
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

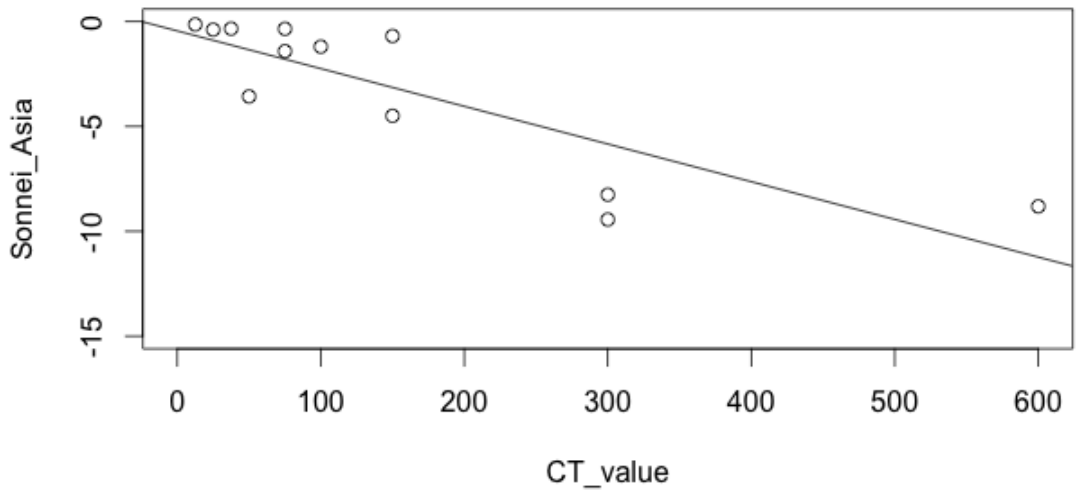
(B)


Flexneri_Asia
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

(C)


Sonnei_Europe
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

(D)


Sonnei_Asia
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

(E)

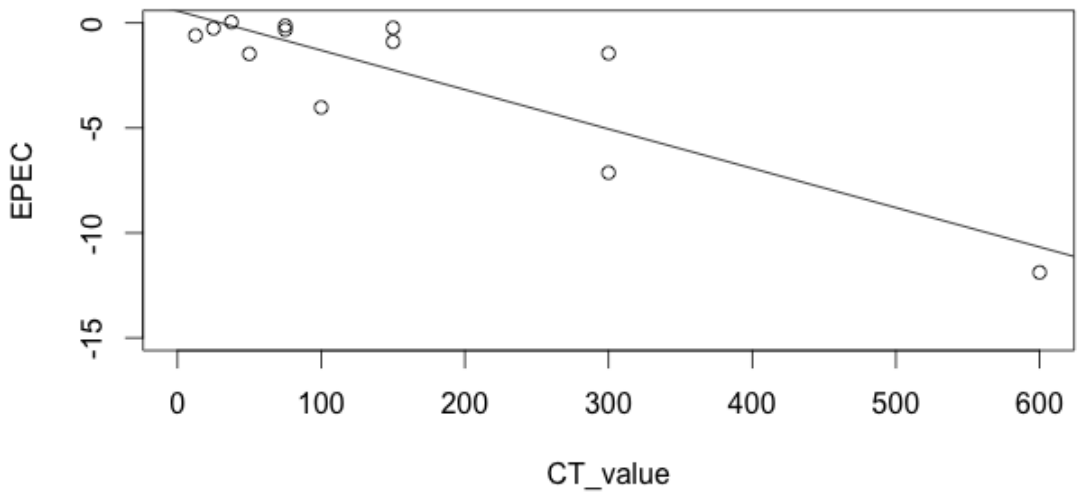


Figure 3. Plots showing the natural log survival fraction (*ln (N/N₀)*) against CT value (mg min L $^{-1}$) for (A) European *S. flexneri,* (B) Vietnamese *S. flexneri,* (C) European *S. sonnei,* (D) Vietnamese *S. sonnei,* and (E) EPEC. All plots show the fitted least squares regression line, with the regression equations given in **Table 2.**

The $CT_{99.9\%}$ values identified from this analysis indicate the dosage of calcium hypochlorite required to reduce bacterial populations of each species group by log^{-3} (**Table 2)**, with *a*, the *y* intercept set at 0. *Shigella sonnei* from Vietnam appears the most resistant to chlorination with this compound, requiring a dose of 383.98 mg min L^{-1} to inactivate the bacteria by 99.9%. For example, with solutions of 20mg/l of calcium hypochlorite, 383.98/20 = 19.20 minutes of contact time would be required on average to reduce bacterial populations of *S. sonnei* by log^{-3}. In contrast, Vietnamese strains of *S. flexneri* had the highest sensitivity to calcium hypochlorite with a $CT_{99.9\%}$ value of 219.22 mg min L^{-1}. In a solution containing 20 mg/l of calcium hypochlorite, 99.9% inactivation would be achieved after 219.22/20 = 10.96 minutes.

Of all *Shigella* strains tested, Vietnamese *S. sonnei* exhibited the greatest resistance to calcium hypochlorite disinfection. There did not appear to be any species level difference in sensitivity between *S. sonnei* and *S. flexneri* when combining strains from Europe and Vietnam (Welch's t-test; t = -0.3865, p-value = 0.7376). This suggests that any difference in resistance to chlorination is attributable to population or strain adaptation and is not a characteristic of the species.

Species group	*Rate constant of disinfection (k) with standard error*	$CT_{99.9\%}$ value (mg min L^{-1})	Regression model *p* value
S. flexneri* (Europe*)	- 0.02056 (S.E. 0.01)	335.9803151	*p = 0.04*
***S. flexneri* (Asia)**	- 0.03151 (S.E. 0.013)	219.2242234	*p = 0.003*
S. sonnei* (Europe*)	- 0.02852 (S.E. 0.008)	242.2074081	*p = 0.01*
***S. sonnei* (Asia)**	- 0.01799 (S.E. 0.014)	383.977503	$p = 1x10^{-4}$
EPEC	- 0.01876 (S.E. 0.006)	368.2172324	*p = 0.015*

Table 2. The Chick-Watson rate constant of disinfection and $CT_{99.9\%}$ values of each species group.

Chlorine sensitivity of Vietnamese Shigella

Modelling the calcium hypochlorite inactivation kinetics of each species group clearly indicates that there is a difference in the average resistance between Vietnamese *S. sonnei* and *S. flexneri*. To determine whether this is due to extreme sensitivity or resistance of a single strain influencing the species average, the inactivation kinetic model was applied to all Vietnamese *Shigella* strains individually. The rate constant of disinfection and $CT_{99.9\%}$ value of each strain in response to calcium hypochlorite inactivation is shown in **Table 3**. There was a significant difference between $CT_{99.9\%}$values of *S. sonnei* and *S. flexneri* (Welch's t test; t = -8.9752, df = 2.373, p-value = 0.0069) with all $CT_{99.9\%}$ values higher in *S. sonnei* isolates, indicating Vietnamese *S. sonnei* strains had a markedly higher chlorination resistance.

Inactivation by sodium dodecyl sulphate (SDS)

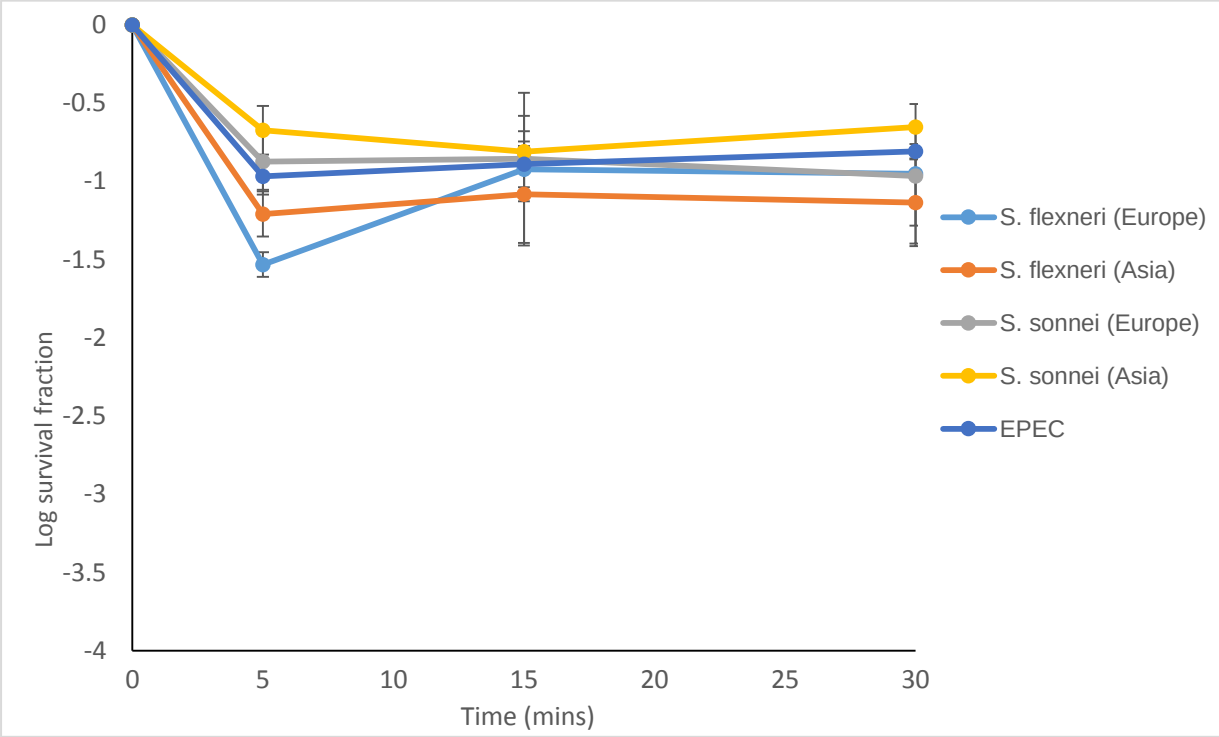
Sensitivity to sodium dodecyl sulphate (SDS) was tested through inactivation assays in solutions of 15, 22.5, and 30 % SDS, with population counts recorded at 0, 5, 15 and 30 minutes. The upper limit of SDS (30%) was chosen as this is the micelle concentration the aqueous reagent[49] and allows for adequate mixing of the test media with the bacterial culture inoculate.

Species and strain	*Rate constant of disinfection (k) with standard error*	$CT_{99.9\%}$ value (mg min L^{-1})	Regression model *p value*
S. flexneri* (MS0052*)	– 0.02998 (S.E. 0.009)	230.435176	$p = 0.006$
***S. flexneri* (DE0350)**	– 0.02875 (S.E. 0.007)	240.253035	$p = 0.027$
S. flexneri* (EG419*)	– 0.03108 (S.E. 0.007)	222.271551	$p = 0.025$
***S. sonnei* (MS004)**	– 0.01936 (S.E. 0.006)	356.805541	$p = 0.009$
S. sonnei* (DE1208*)	– 0.01676 (S.E. 0.008)	412.231024	$p = 0.05$
***S. sonnei* (EG0430)**	– 0.01721 (S.E. 0.005)	401.31036	$p = 0.01$

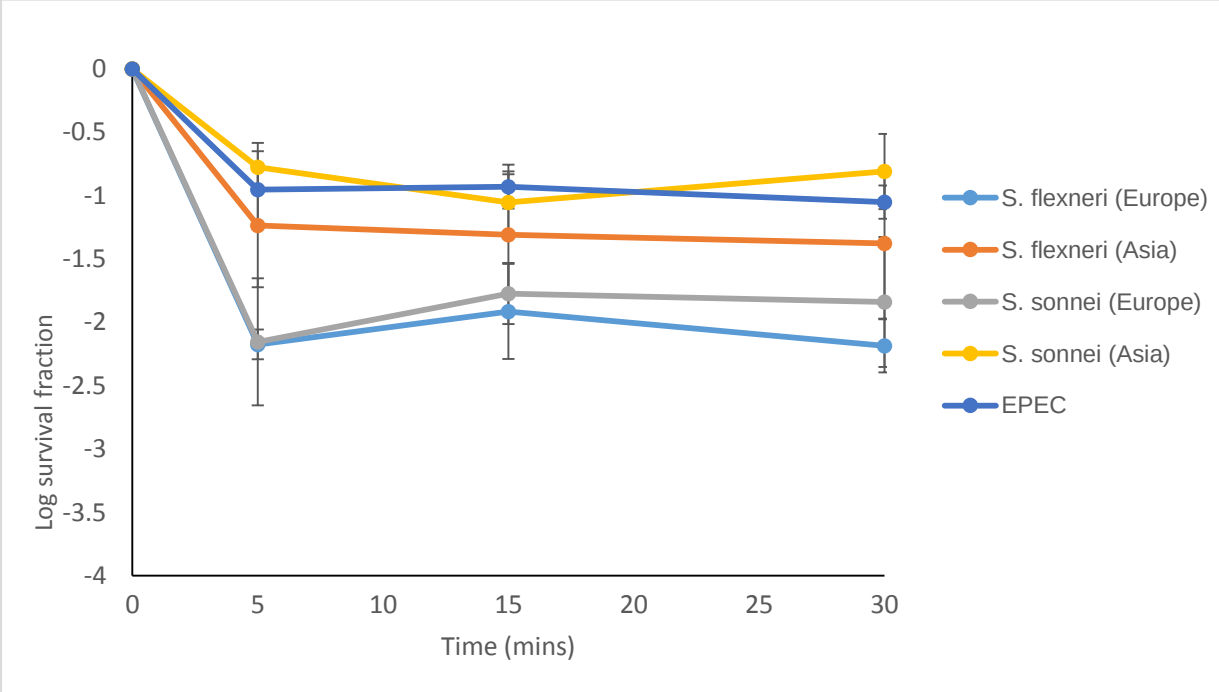
Table 3. The Chick-Watson rate constant of disinfection and $CT_{99.9\%}$ values of Vietnamese *Shigella* strains.

The log survival fraction of all species groups in varying concentrations of SDS is illustrated in **Figure 4.** The response of all species groups to SDS is characterised by a rapid decrease in the log population size after 5 mins of disinfection followed by a plateau or slight increase until the termination of the assay at 30 mins. The magnitude of the initial population decrease is determined by the concentration of SDS in solution, with 15% SDS yielding a change in survival fraction of ~ $1x10^{-1}$ ($\pm 0.5x10^{1}$) and ~$1x10^{-2}$ ($\pm 1x10^{2}$) in 30% SDS. This pattern of bacterial survival can be explained by two alternative theories. Firstly, it may be that SDS is acting as a weak disinfectant on the bacteria, where only the susceptible organisms are killed before the populations stabilise. Alternatively, there may have been the initial population decrease due to a lowering of the carrying capacity of the nutrient media in solutions as the concentration of SDS increases. In this situation SDS would not be having a direct antibacterial effect and the maximum load of the bacteria will be defined by the available nutrients. If the reduction in bacterial survival is indeed due to a weak antimicrobial action of the detergent, the maximum concentration of SDS in cleaning products is 30%[50], and thus the levels needed to achieve effective killing of *Shigella* and *E. coli* groups in isolation was higher than is in commercial use.

(A)


0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Log survival fraction
0
5
10
15
20
25
30
Time (mins)
S. flexneri (Europe)
S. flexneri (Asia)
S. sonnei (Europe)
S. sonnei (Asia)
EPEC

(B)


0
-0.5
-1
-1.5
-2
-2.5
-3
-3.5
-4
Log survival fraction
0
5
10
15
20
25
30
Time (mins)
S. flexneri (Europe)
S. flexneri (Asia)
S. sonnei (Europe)
S. sonnei (Asia)
EPEC

(C)

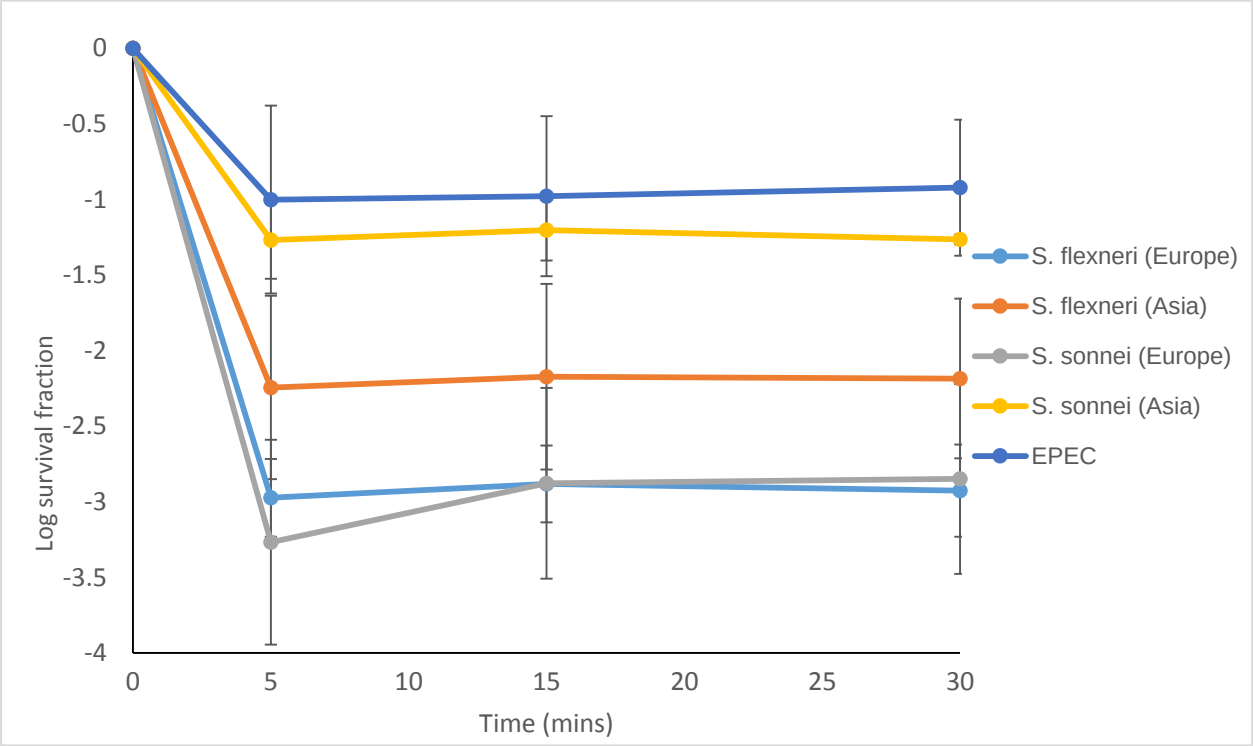


Figure 4. Survival of *Shigella* species and *EPEC* against contact time in minutes in SDS solutions. The log survival fraction (N/N_o) is shown for each tested concentration of SDS, (A) 15%, (B) 22.5%, and (C) 30%. Error bars denote the standard error at each recorded point.

Competitive fitness of Vietnamese S. sonnei and S. flexneri

The relative competitive fitness of Vietnamese *Shigella flexneri* and *S. sonnei* strains in disinfectant solutions was investigated by conducting pairwise competition assays with all isolates of *S. sonnei* against all *S. flexneri* in varying sub-lethal concentrations of calcium hypochlorite (0, 1.25, 2.5, and 5 mg/l) for 24 hours. The initial and final population counts were taken in these experiments rather than a series of time points as the disinfectant concentration will decline within this timeframe. The rationale behind these experiments is to determine the extent that *S. sonnei* and *S. flexneri* will respond to disinfectant injury when in direct competition. As reported previously, SDS solutions at these concentrations did not have any significant effect on the survival of Vietnamese *Shigella* isolates in inactivation experiments over longer contact times, and thus these competition assays were not investigated further (data is available in **Supplementary materials).**

The relative fitness of two strains in competition can be expressed as the difference between the Malthusian (exponential) growth of each strain at time point, *t*. The Malthusian growth coefficient of each population, m_{pop}, can be expressed as the natural log of the colony count of each population at time *t* over the initial population count:

$$m_{pop} = ln\left(\frac{N_t}{N_o}\right)$$

In growing populations, the fitness coefficient, *w,* will be the ratio between the Malthusian growth of each population (m_a and m_b) [51]:

$$w = m_a / m_b$$

When population growth is equal, the fitness coefficient, *w*, will give a dimensionless value of 1. Any deviation will suggest a difference in growth rates, i.e. *w* = 0.5 would indicate that the growth rate of m_b is 50% faster than m_a, and *w* = 1.5 the reverse.

Competitive fitness in calcium hypochlorite

In the population count data from these experiments, though, I observed some treatments where one or more of the populations in direct competition were in decline and thus the fitness coefficient, *w,* would not be suitable for these data. To account for instances where one or both populations may be in decline, the selection rate, *r,* was instead applied, which is the difference between the Malthusian growth coefficients of each population at time *t* [52]:

$$r = \frac{(m_a - m_b)}{t}$$

Concentration of calcium hypochlorite	Selection coefficient (*r*)			
	0 mg/l	**1.25 mg/l**	**2.5 mg/l**	**5 mg/l**
S. sonnei MS / S. flexneri MS	1.322299397	1.873962467	7.177348345	13.92256376
S. sonnei MS / S. flexneri DE	1.313153982	4.286032651	11.48572123	10.59724976
S. sonnei MS / S. flexneri EG	-0.064143801	0.591695536	13.55299336	20.38761824
S. sonnei DE / S. flexneri MS	1.58532322	3.874333341	9.560730874	24.4261022
S. sonnei DE / S. flexneri DE	-0.712741406	5.578419462	9.788150591	33.21928095
S. sonnei DE / S. flexneri EG	0.877691624	1.084635482	1.228615623	16.69518288
S. sonnei EG / S. flexneri MS	-0.23317226	0.428843299	6.760084964	23.22419103
S. sonnei EG / S. flexneri DE	0.204951973	2.56130666	15.03511972	33.90598993
S. sonnei EG / S. flexneri EG	0.867848033	-0.799887462	2.162574524	11.36861478

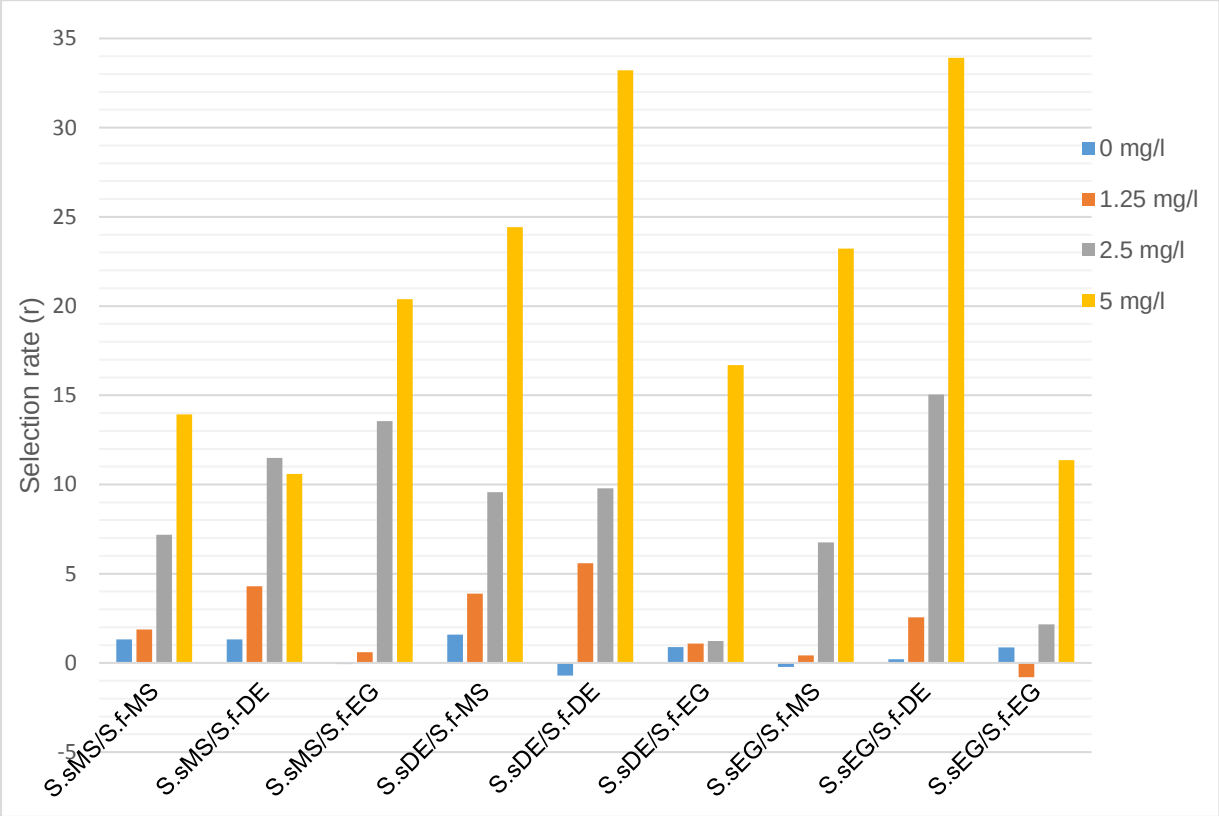


Figure 5. The selection coefficients (*r*) of competition assays from all-against-all experiments of Vietnamese *S. sonnei* and *S. flexneri* at varying concentrations of calcium hypochlorite.

When the both populations are in growth or decline equally, $r = 0$. A positive r value will indicate that the Malthusian growth rate, m_a, is greater than m_b, measured in t^{-1}. For example, when $r = 5$ and $t = 24$ hours, the Malthusian rate m_a will have increased by 5 natural logs, or e^5, more than m_b in 24 hours.

Excluding a single pairwise assay (*S. sonnei* EG0430 vs *S. flexneri* EG419; 1.25mg/l calcium hypochlorite), all Vietnamese *S. sonnei* had markedly higher Malthusian growth rates than *S. flexneri* when in competition at all concentrations of calcium hypochlorite (**Figure 5)**. The difference in growth rates were most striking at higher concentrations (>2.5mg/l), where all *S. sonnei* strains had a significantly higher growth rate than *S. flexneri,* with some occurrences of declining population sizes to undetectable levels (<10 CFU/ml) in the latter species (**Supplementary materials).** At concentrations of 5 mg/l, the difference in growth rates was the highest, with *S. sonnei* having an advantage of between ~ 10.5 – 34 natural log growth in 24 hours over *S. flexneri* strains.

It is also interesting to note that a concentration of 5 mg/l calcium hypochlorite in solution was within a resistance range for both Vietnamese *S. sonnei* and *S. flexneri* after 30 minutes of contact time (**Figure 2B)**. In these competition experiments, though, the populations of *S. flexneri* declined on average in seven of the nine pairwise assays, with bacterial populations undetectable in *S. flexneri* DE0350 in two assays (**Supplementary materials).** *S. sonnei* populations increased in all treatments, which suggests that *S. sonnei* was having an antagonistic effect on *S. flexneri* growth in the presence of high concentrations of calcium hypochlorite.

2.5 Discussion

The replacement of *Shigella flexneri* as the primary cause of bacillary dysentery by *S. sonnei* in many developing countries[2–4] has impacted the management and treatment of the disease in these regions. Understanding the underlying reasons for the successful emergence of *S. sonnei* is important to develop strategies to stop the spread of infection. One proposed hypothesis is that the increasing use of cleaning products and improving sanitation may be contributing to a shift in the dominant species of *Shigella* in these regions[39,52], and this may be driven, in part, by the species' ability to withstand disinfection of environmental sources of transmission. To my knowledge, this study is the first to directly compare chlorine and detergent

disinfection in *Shigella* species using clinical isolates from populations undergoing this species replacement.

Though it is difficult to accurately predict the absolute efficacy of disinfectants on these bacterial pathogens due to the limited number of samples tested in this study, the results of the work presented here suggest that there are significant differences in resistance to calcium hypochlorite disinfection between Vietnamese *S. sonnei* and *S. flexneri* isolates in the tested strains. This has been demonstrated through both inactivation assays and pairwise competitive growth experiments. There was also validation of the antibacterial potential of chlorine disinfection on *Shigella* and enteropathogenic *E. coli*, with all populations reduced by 99.9% after 30 mins contact time in solutions of 20 mg/l calcium hypochlorite.

The evidence for any direct antimicrobial action by SDS on both *Shigella* and EPEC was less clear. The response of all populations to SDS treatment was characterised by a rapid decrease in the surviving fraction of populations followed by a plateau, with the magnitude of the decrease in population size from the initial count dependent on the percentage concentration of SDS in solution **(Figure 4)**. The reaction of populations in this way is described as tolerance to a particular compound, where an organism is able to survive transient exposure to high concentrations of a treatment [53]. Alternatively, this period of microbial death after inoculation could be attributed to the lower nutrient load available for viable growth in high SDS concentration solutions, lowering the carrying capacity for supporting bacterial growth. This may explain the stabilizing of the population size after initial contact with the detergent. Previously there has only been limited evidence for SDS to act as an antibacterial, which was linked to biofilm formation in very high concentrations of the detergent[33]. The results of this study do not lend conclusive support for direct biocidal action by SDS on *Shigella* species or EPEC.

All strains of *Shigella* and EPEC decreased in population size when tested against calcium hypochlorite above 2.5mg/l in solution. When considering the difference between Vietnamese populations in sensitivity to this treatment, the levels of calcium hypochlorite required for 99.9% inactivation of bacterial populations as a product of concentration and contact time (the $CT_{99.9\%}$ value), Vietnamese *S. sonnei* were able to withstand significantly higher concentrations of disinfectant (356.80 – 412.23 mg min L^{-1}) than *S. flexneri* (222.27 – 240.25 mg min L^{-1}) (**Table 3)**. These figures should not be taken as the absolute quantity of calcium hypochlorite required to inactivate the organisms to these levels as it has been shown that using a higher

concentration with a lower contact time will have a greater biocidal impact[14]. For a comparison of the average resistance between strains though, this measure is effective for discerning variation.

When co-cultured in direct competition with solutions containing calcium hypochlorite, Vietnamese *S. sonnei* strains possessed a greater selective advantage over *S. flexneri*, with the difference in the population survival fraction increasing with disinfectant concentration. There was also evidence that *S. sonnei* can antagonistically affect the survival of *S. flexneri* by increasing its susceptibility to calcium hypochlorite, which causes a significant reduction in some *S. flexneri* bacteria if in solutions that were within the tolerable range for the strains when grown separately (**Figure 5)**. There are different mechanisms by which a species can inhibit the growth of another when in mixed cultures. There may be competition for metabolic substrates, a change to the environment through toxins released by one species, or a build-up of metabolic waste[54,55]. It may also be that calcium hypochlorite at these lower concentrations is causing sub-lethal injury to *S. flexneri* bacteria rather than killing the cells but populations are unable to respond in mixed cultures where the stress on microorganisms will be greater. It would be interesting to further investigate the extent to which *Shigella* species interact to influence the sensitivity to various disinfectants.

Previously*, S. sonnei* has been reported to be readily taken up and maintained by the common environmental amoeba, *Acanthamoeba,* where the encapsulated microorganism will be protected from extreme environmental conditions[56,57]. *S. flexneri,* though, will inhibit the growth of *Acanthamoeba* through the induction of apoptosis in the amoeba cell[58,59]. It has been hypothesised that *S. sonnei* may have an ability to withstand greater levels of chlorination than *S. flexneri* due to the protection offered from this phagocytic process and this may be contributing to the replacement of *S. flexneri* with *S. sonnei* in the developing world[39]. The experimental results here, however, indicate that there are significant differences in chlorine resistance between *Shigella* strains isolated from a region undergoing species replacement, even in the absence of *Acanthamoeba*. This suggests that there are key intrinsic differences between the species that influences their sensitivity to chlorine disinfection.

Inactivation experiments with European isolates of *S. sonnei* and *S. flexneri* suggested that the differences in chlorine sensitivity identified in Vietnamese isolates does not extend to the species-level. Combined analysis of all *Shigella* strains found

no meaningful difference in resistance between *S. sonnei* and *S. flexneri* (t = -0.3865, p = 0.7376), and there was also no distinction between the $CT_{99.9\%}$ values when considering only the European *Shigella* isolates (t = 2.4872, p = 0.1243). In addition, the EPEC strains tested showed high levels of resistance to chlorination, falling between that of the least sensitive group, Vietnamese *S. sonnei,* and European *S. flexneri*. *S. flexneri* is known to share many chromosomal genes with pathogenic and non-pathogenic *E. coli*[60], whereas *S. sonnei* has been proposed as one the most divergent serotypes[61,62]. This suggests that *S. sonnei* in Vietnam may have undergone some local adaptation that has increased their resistance to chlorine in response to the selective pressure of disinfection. These phenotypic alterations may be associated with the process by which chlorine will disrupt and kill bacterial cells.

While *Shigella* strains used in this study were selected to represent the major lineages and serotypes found in Vietnam, a limitation with this work is that the full range of global genetic variation found in each species in either the Vietnamese or European samples. It may be, therefore, suggested that the results presented here are more reflective of lineage specific characteristics rather than species or geographic variation. European strains were selected predominately for the availability of sequencing data, again not representing the full evolutionary range of *Shigella* strains present globally. There was some heterogeneity of samples in each species, with one *S. sonnei* lineage 1 and one *S. flexneri* serotype 1a sample in the European strains, though we did not find any differences in chlorine resistance between these strains compared to other European strains of the same species (**Supplementary data)**.

The proposed mechanism for the chlorine inactivation of microorganism populations is by attacking the cellular envelope of the cell and compromising the integrity of cell wall[16,24,25]. This will lead to interactions with components of the cell membrane that can affect the permeability of the cell across channels in the membrane. Two recent studies have proposed that efflux pumps may have an impact on chlorine resistance in bacteria, with the effect of the disinfectant also linked to antibiotic resistance[63,64]. The role of efflux pumps in resistance to antibiotics and bacterial fitness has been studied in detail, with the extrusion of drugs and toxic substances found to be an significant mechanism for maintaining homeostasis in viable cells[65–68]. Evaluating the importance of membrane pumps in chlorine resistance has motivated the research presented in **Chapter 3** of this thesis.

In conclusion, inactivation and competition assays investigating the sensitivity of clinical isolates of *Shigella sonnei* and *Shigella flexneri* in Vietnam to calcium hypochlorite and SDS detergent disinfection found significant differences between the species' ability to survive chlorination, though SDS did not have a direct antimicrobial effect. Comparing survival of these strains against European *S. sonnei* and *S. flexneri* strains, as well as European enteropathogenic *E. coli*, indicated that this variation in chlorine sensitivity is likely due to local differences in the populations and is not characteristic of each species. Though the number of isolates tested in this study is limited, the results presented here suggest further work to determine the impact of disinfectant resistance and the mechanisms by which bacteria become resistant can be important for managing microbial diseases.

References

1. Kotloff, K. L. *et al.* Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull. World Health Organ.* **77,** 651–66 (1999).

2. Banga Singh, K.-K., Ojha, S. C., Deris, Z. Z. & Rahman, R. A. A 9-year study of shigellosis in Northeast Malaysia: Antimicrobial susceptibility and shifting species dominance. *Z. Gesundh. Wiss.* **19,** 231–236 (2011).

3. Qu, F. *et al.* Genotypes and antimicrobial profiles of Shigella sonnei isolates from diarrheal patients circulating in Beijing between 2002 and 2007. *Diagn. Microbiol. Infect. Dis.* **74,** 166–70 (2012).

4. Vinh, H. *et al.* A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation. *BMC Infect. Dis.* **9,** 204 (2009).

5. García-Fulgueiras, A. *et al.* A large outbreak of Shigella sonnei gastroenteritis associated with consumption of fresh pasteurised milk cheese. *Eur. J. Epidemiol.* **17,** 533–538 (2001).

6. Bagamboula, C. F., Uyttendaele, M. & Debevere, J. Growth and survival of Shigella sonnei and S . flexneri in minimal processed vegetables packed under equilibrium modified atmosphere and stored at 7 1 C and 12 1 C. *Food Microbiol.* **19,** 529–536 (2002).

7. Karasz, O'Reilly & Bair. © 1964 Nature Publishing Group. *Nature* **202,** 693–694 (1964).

8. Faruque, S. M. *et al.* Isolation of Shigella dysenteriae Type 1 and S . flexneri Strains from Surface Waters in Bangladesh : Comparative Molecular Analysis of Environmental Shigella Isolates versus Clinical Strains Isolation of Shigella dysenteriae Type 1 and S . flexneri Strai. *Appl. Environ. Microbiol.* **68,** 3908–3913 (2002).

9. Islam, M. S., Hasan, M. K. & Khan, S. I. Growth and survival of Shigella flexneri in common Bangladeshi foods under various conditions of time and temperature. *Appl. Environ. Microbiol.* **59,** 652–654 (1993).

10. Colwell, R. R. Survival of S Dysenteriae Type 1 on Fomites.Pdf. **19,** 177–182 (2016).

11. Reller, M. E. *et al.* A large, multiple-restaurant outbreak of infection with Shigella flexneri serotype 2a traced to tomatoes. *Clin. Infect. Dis.* **42,** 163–169 (2006).

12. Mermel, L. a *et al.* Outbreak of Shigella sonnei in a clinical microbiology laboratory. *J. Clin. Microbiol.* **35,** 3163–5 (1997).

13. Uren NG, Crake T, L. D. The New England Journal of Medicine as published by New England Journal of Medicine. Downloaded from www.nejm.org on July 28, 2010. For personal use only. No other uses without permission. Copyright © 1994 Massachusetts Medical Society. All rights reserve. (1994).

14. Huang, J. J. *et al.* Inactivation and reactivation of antibiotic-resistant bacteria by chlorination in secondary effluents of a municipal wastewater treatment plant. *Water*

Res. **45,** 2775–2781 (2011).

15. Rice, E. W., Clark, R. M. & Johnson, C. H. Chlorine Inactivation of Escherichia coli O157 : H7. **5,** 461–463 (1999).

16. Virto, R., Mañas, P., Álvarez, I., Condon, S. & Raso, J. Membrane damage and microbial inactivation by chlorine in the absence and presence of a chlorine-demanding substrate. *Appl. Environ. Microbiol.* **71,** 5022–5028 (2005).

17. LeChevallier, M. W., Singh, A., Schiemann, D. A. & McFeters, G. A. Changes in virulence of waterborne enteropathogens with chlorine injury. *Appl. Environ. Microbiol.* **50,** 412–419 (1985).

18. LeChevallier, M. W., Cawthon, C. D. & Lee, R. G. Inactivation of biofilm bacteria. *Appl. Environ. Microbiol.* **54,** 2492–2499 (1988).

19. Lechevallier, M. W. *et al.* Factors promoting survival of bacteria in chlorinated water supplies . Factors Promoting Survival of Bacteria in Chlorinated Water Supplies. **54,** 649–654 (1988).

20. Kotula, K. L. (University of D. N. ., Kotula, A. W., Rose, B. E., Pierson, C. J. & Camp, M. Reduction of aqueous chlorine by organic material. *J. food Prot.* (1997).

21. Dukan, S. & Toutati, D. Hypochlorous acid stress in Escherichia coli : resistance , DNA damage , and comparison with hydrogen peroxide stress . Hypochlorous Acid Stress in Escherichia coli : Resistance , DNA Damage , and Comparison with Hydrogen Peroxide Stress. *J. Bacteriol.* **178,** 6145–6150 (1996).

22. Buchholz, A. & Matthews, K. R. Reduction of Salmonella on alfalfa seeds using peroxyacetic acid and a commercial seed washer is as effective as treatment with 20 000 ppm of Ca (OCl) 2. **51,** 462–468 (2010).

23. Dumani, A. *et al.* Antibacterial Efficacy of Calcium Hypochlorite with Vibringe Sonic Irrigation System on Enterococcus faecalis : An In Vitro Study. **2016,** (2016).

24. Venkobachar, C., Iyengar, L. & Prabhakara Rao, A. V. S. Mechanism of disinfection: Effect of chlorine on cell membrane functions. *Water Res.* **11,** 727–729 (1977).

25. Benarde, M. a., Snow, W. B., Olivieri, V. P. & Davidson, B. Kinetics and mechanism of bacterial disinfection by chlorine dioxide. *Appl. Microbiol.* **15,** 257–65 (1967).

26. Mir, J., Morató, J. & Ribas, F. Resistance to chlorine of freshwater bacterial strains. *J. Appl. Microbiol.* **82,** 7–18 (1997).

27. Barrette, W. C., Hannum, D. M., Wheeler, W. D. & Hurst, J. K. General mechanism for the bacterial toxicity of hypochlorous acid: abolition of ATP production. *Biochemistry* **28,** 9172–9178 (1989).

28. Bernofsky, C. Nucleotide chloramines and neutrophil-mediated cytotoxicity. *FASEB J. Off. Publ. Fed. Am. Soc. Exp. Biol.* **5,** 295–300 (1991).

29. Dennis, W. H., Olivieri, V. P. & Krus??, C. W. The reaction of nucleotides with aqueous hypochlorous acid. *Water Res.* **13,** 357–362 (1979).

30. Sirisattha, S., Momose, Y., Kitagawa, E. & Iwahashi, H. Toxicity of anionic detergents

determined by Saccharomyces cerevisiae microarray analysis. *Water Res.* **38,** 61–70 (2004).

31. Hosseini, F., Malekzadeh, F., Amirmozafari, N. & Ghaemi, N. Biodegradation of anionic surfactants by isolated bacteria from activated sludge. *Int. J. Environ. Sci. Technol.* **4,** 127–132 (2007).

32. Paulo, A. M. S., Plugge, C. M., García-Encina, P. a. & Stams, A. J. M. Anaerobic degradation of sodium dodecyl sulfate (SDS) by denitrifying bacteria. *Int. Biodeterior. Biodegradation* **84,** 14–20 (2013).

33. Simões, M., Simões, L. C., Pereira, M. O. & Vieira, M. J. Sodium dodecyl sulfate allows the persistence and recovery of biofilms of Pseudomonas fluorescens formed under different hydrodynamic conditions. *Biofouling* **24,** 35–44 (2008).

34. Nickerson, K. W. & Aspedon, A. MicroReview Detergent-shock response in enteric bacteria. **6,** 957–961 (1992).

35. Adamowicz, M., Kelley, P. M. & Nickerson, K. W. Detergent (sodium dodecyl sulfate) shock proteins in Escherichia coli. *J. Bacteriol.* **173,** 229–233 (1991).

36. Hofmann, A. F. & Eckmann, L. How bile acids confer gut mucosal protection against bacteria. *Pnas* **103,** 4333–4334 (2006).

37. Trasande, L. *et al.* How developing nations can protect children from hazardous chemical exposures while sustaining economic growth. *Health Aff.* **30,** 2400–2409 (2011).

38. Zhang, W. D. & DiGiano, F. A. Comparison of bacterial regrowth in distribution systems using free chlorine and chloramine: a statistical study of causative factors. *Water Res.* **36,** 1469–1482 (2002).

39. Thompson, C. N., Duy, P. T. & Baker, S. The Rising Dominance of Shigella sonnei: An Intercontinental Shift in the Etiology of Bacillary Dysentery. *PLoS Negl. Trop. Dis.* **9,** e0003708 (2015).

40. Ahmed, Z. U., Sarker, M. R. & Sack, D. a. Nutritional requirements of shigellae for growth in a minimal medium. *Infect. Immun.* **56,** 1007–9 (1988).

41. Bh, C. *et al.* A Shigella flexneri Virulence Plasmid Encoded Factor Controls Production of Outer Membrane Vesicles. **4,** 2493–2503 (2014).

42. e Silva Leonardo, N. G. *et al.* Calcium Hypochlorite Solutions: Evaluation of Surface Tension and Effect of Different Storage Conditions and Time Periods over pH and Available Chlorine Content. *J. Endod.* **42,** 641–645 (2016).

43. Ito, H., Kido, N. & Kato, N. Possible Mechanisms Underlying the Slow Lactose Fermentation Phenotype in Shigella spp . **57,** 2912–2917 (1991).

44. R Development Core Team, R. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **1,** 409 (2011).

45. Chick, H. An Investigation of the Laws of Disinfection. *J. Hyg. (Lond).* **8,** 92–158 (1908).

46. Watson, H. E. A Note on the Variation of the Rate of Disinfection with Change in the Concentration of the Disinfectant. *J. Hyg. (Lond).* **8,** 536–542 (1908).

47. Le Dantec, C. *et al.* Chlorine disinfection of atypical mycobacteria isolated from a water distribution system. *Appl. Environ. Microbiol.* **68,** 1025–1032 (2002).

48. Bales, B. L., Messina, L., Vidal, A., Peric, M. & Nascimento, O. R. Precision Relative Aggregation Number Determinations of SDS Micelles Using a Spin Probe. A Model of Micelle Surface Hydration. *J. Phys. Chem. B* **102,** 10347–10358 (1998).

49. Bondi, C. A. M. *et al.* Human and Environmental Toxicity of Sodium Lauryl Sulfate (SLS): Evidence for Safe Use in Household Cleaning Products. 27–32 (2015). doi:10.4137/EHI.S31765.TYPE

50. Lenski, R. E., Rose, M. R., Simpson, S. C. & Tadler, S. C. Long-Term Experimental Evolution in Escherichia coli. I. Adaptation and Divergence During 2,000 Generations. *Am. Nat.* **138,** 1315 (1991).

51. Applebee, M. K., Herrgård, M. J. & Palsson, B. Ø. Impact of individual mutations on increased fitness in adaptively evolved strains of Escherichia coli. *J. Bacteriol.* **190,** 5087–5094 (2008).

52. Holt, K. E. *et al.* Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44,** 1056–9 (2012).

53. Brauner, A., Fridman, O., Gefen, O. & Balaban, N. Q. Distinguishing between resistance, tolerance and persistence to antibiotic treatment. *Nat. Rev. Microbiol.* **14,** 320–30 (2016).

54. Schiemann, D. A. & Olson, S. A. Antagonism by gram-negative bacteria to growth of Yersinia enterocolitica in mixed cultures. *Appl. Environ. Microbiol.* **48,** 539–544 (1984).

55. Long, R. A., Eveillard, D., Franco, S. L. M., Reeves, E. & Pinckney, J. L. Antagonistic interactions between heterotrophic bacteria as a potential regulator of community structure of hypersaline microbial mats. **83,** 74–81 (2013).

56. Jeong, H. J. *et al.* Acanthamoeba: Could it be an environmental host of Shigella? *Exp. Parasitol.* **115,** 181–186 (2007).

57. Saeed, A., Abd, H., Edvinsson, B. & Sandström, G. Acanthamoeba castellanii an environmental host for Shigella dysenteriae and Shigella sonnei. *Arch. Microbiol.* **191,** 83–88 (2008).

58. Saeed, A., Johansson, D., Sandström, G. & Abd, H. Temperature Depended Role of Shigella flexneri Invasion Plasmid on the Interaction with Acanthamoeba castellanii. *Int. J. Microbiol.* **2012,** 917031 (2012).

59. Zychlinsky, A. *et al.* In vivo apoptosis in Shigella flexneri infections. *Infect. Immun.* **64,** 5357–5365 (1996).

60. Yang, F. *et al.* Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.* **33,** 6445–58 (2005).

61. Pupo, G. M., Lan, R. & Reeves, P. R. 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 (2000).

62. Escobar-Páramo, P., Giudicelli, C., Parsot, C. & Denamur, E. The evolutionary history of Shigella and enteroinvasive Escherichia coli revised. *J. Mol. Evol.* **57,** 140–148 (2003).

63. Yuan, Q. Bin, Guo, M. T. & Yang, J. Fate of antibiotic resistant bacteria and genes during wastewater chlorination: Implication for antibiotic resistance control. *PLoS One* **10,** 1–11 (2015).

64. Prasad Karumathil, D., Yin, H. B., Kollanoor-Johny, A. & Venkitanarayanan, K. Effect of chlorine exposure on the survival and antibiotic gene expression of multidrug resistant Acinetobacter baumannii in water. *Int. J. Environ. Res. Public Health* **11,** 1844–1854 (2014).

65. Hirakawa, H. *et al.* AcrS/EnvR represses expression of the acrAB multidrug efflux genes in Escherichia coli. *J. Bacteriol.* **190,** 6276–6279 (2008).

66. Marcusson, L. L., Frimodt-Møller, N. & Hughes, D. Interplay in the selection of fluoroquinolone resistance and bacterial fitness. *PLoS Pathog.* **5,** e1000541 (2009).

67. Kugelberg, E., Löfmark, S., Wretlind, B. & Andersson, D. I. Reduction of the fitness burden of quinolone resistance in Pseudomonas aeruginosa. *J. Antimicrob. Chemother.* **55,** 22–30 (2005).

68. Ruiz, J. Mechanisms of resistance to quinolones: target alterations, decreased accumulation and DNA gyrase protection. *J. Antimicrob. Chemother.* **51,** 1109–17 (2003).

Chapter 3

The role of efflux pumps in chlorine resistance in *Shigella sonnei* and *Shigella flexneri*

3.1 Abstract

Transport proteins situated in the cellular envelope of bacteria, termed efflux pumps, have been shown to play an important role in removing toxic substances from within the cell. A number of these pumps can extrude a variety of substrates, including antibiotics, detergents and metals. Importantly, it has been shown that overexpression of the tripartite efflux system AcrAB-TolC, found in *E. coli* and *Shigella,* is involved in conferring fluoroquinolone resistance. Furthermore, efflux pumps have been proposed to have a direct association with resistance to chlorine disinfection in Gram-negative bacteria.

The aim of this study is to determine whether these generalised efflux pumps will be involved in the response to chlorine disinfection in *Shigella* species, and in particular, whether any differences in resistance to chlorine can be explained by variation in genes encoding and regulating these pumps. In addition, resistance to fluoroquinolones is characterised in the *Shigella* strains, with the purpose of finding links between resistance to these antimicrobials and different levels of chlorine resistance.

The results of this chapter show that, overall, there is significant decrease in bacterial survival in the presence of chlorine when efflux pumps are inhibited by carbonyl cyanide 3-chlorophenylhydrazone (CCCP). Unfortunately, due to the small number of fluoroquinolone resistant strains in this study I was unable to robustly discern any differences in chlorine resistance between drug resistant and susceptible isolates. Finally, though there was variation detected in genes associated with efflux pumps, no candidates for increased resistance could be confidently identified without further work on differential expression or functional changes.

These findings lend support to previous studies finding that efflux pumps play an important role in the response to chlorination in bacteria, particularly at high concentrations, with this work the first to describe this process in *Shigella.* Though I was unable to detect any key mutational differences that correlated with higher chlorine resistance, further work looking at expression levels in pump encoding genes could help to better understand the bacterial response to chlorination, and the development of disinfectant resistance.

3.2 Introduction

Both Gram–positive and Gram–negative bacteria contain transport proteins within the cellular membrane, known as efflux pumps, that are involved in the removal of substrates, including toxic substances, out of the cell[1,2]. Pumps can be substrate-specific, such as the tetracycline-specific efflux pump encoded by the *tetA* gene[3], though many will be able to move a range of different compounds, including multiple classes of antibiotics[4–6] and other potentially harmful materials, such as heavy metals[7]. In addition, there is evidence that these proteins can be important for other physiological processes in pathogens including increasing virulence by moving host-defence molecules out from within the cell[8].

Bacterial efflux pumps are divided into five main categories that can be composed of either a single channel or made up of multiple sections that can span both the inner and outer membrane depending on the taxon[9]. The five families of efflux pumps are; resistance-nodulation-division (RND), major facilitator superfamily (MFS), ATP-binding cassette superfamily (ABC), small multidrug resistance (SMR), and multidrug and toxic compound extrusion (MATE). The most common mechanism by which these proteins move compounds out of cells is by utilising a proton-motive force generated by an electron transport chain across the membrane[10].

The RND family of pumps in particular have been shown to be important in antibiotic resistance in Gram-negative bacteria, for example, the AcrAB-TolC pump in *Escherichia coli* that is able to transport a variety of dissimilar compounds and, as such, is known as a multidrug efflux pump. The transporter protein will lower the intracellular concentration of antibiotic through active expulsion of the toxin, thus increasing the minimum inhibitory concentration (MIC) of the drug required to kill the organism. In prokaryotes, genes encoding these proteins can be found on both the chromosomal genome and on plasmids[9], though those controlled by genes on the chromosome will be of most importance as these will relate to innate mechanisms in bacteria, including taxa with low levels of recombination. An over-expression in pump genes has been shown to be associated with resistance to antimicrobials[1,5,6,11], as well as an increase in the movement of other substances across the membrane, such as dyes and detergents[12].

In chapter 2 of this thesis I discussed the differential survival rates discovered in Vietnamese *Shigella sonnei* and *S. flexneri* strains when subjected to chlorine disinfection and how this may relate to the successful emergence of *S. sonnei* strains in developing regions. It has been proposed that the same efflux pumps that are involved in resistance to antimicrobials can also affect bacterial sensitivity to disinfectants, including chlorine[13,14]. This non-specificity of drug efflux systems suggests the presence of greater resistance to one selective pressure

can automatically confer a level of cross-resistance against other toxic compounds[1], and thus it can be theorised that antibiotic resistant bacteria will also have a reduced sensitivity to chlorine. In addition, exposure to chlorine has been proposed to increase susceptibility to some antibiotics[13–15], suggesting an association in the response to these toxic substances.

To investigate whether efflux pumps are involved in the bacterial response to chlorination and potential resistance in *Shigella,* I have conducted *in vitro* inactivation assays on six strains of each *Shigella* species (three Vietnamese and three of European origin), both in the presence and absence of a general efflux pump inhibitor, carbonyl cyanide 3-chlorophenylhydrazone (CCCP). This compound will inactivate transport pumps by disrupting the gradient potential and has been shown to increase the susceptibility of resistant bacteria to antibiotics[16,17].

In addition, whole genome sequence data is available for the *Shigella* isolates used in this study and thus it is possible to look for key variants that are likely to confer antibiotic resistance to fluoroquinolones. There has been extensive research on the role of efflux pumps in resistance to quinolones in Gram-negative bacteria, particularly overexpression in genes of the AcrAB-TolC pump located in *E. coli* and *Shigella* cellular envelopes[11,18–20], and thus I will also be looking for significant correlations between drug and chlorine resistance. Finally, mutations in pump genes will be identified through the analysis of variation in the chromosomal genomes of the tested strains within a set of genes previously described as being associated with efflux pumps. The aim is to determine whether variation in these genes will be associated with differential susceptibility to disinfectants, as well as any links to fluoroquinolone resistance.

3.3 Materials and methods

Bacterial isolates, sample preparation and population counts

Details of the *Shigella* isolates used in this study are available with accompanying metadata in chapter 2 (**Chapter 2, Table 1).** Live bacterial cultures were stored and prepared for inoculation, and population counts estimated in CFU/ml, using the methods stated in chapter 2.

Inactivation assays

The role of efflux pumps in *Shigella* chlorine disinfection was examined by comparing the survival of bacterial populations in varying concentrations of calcium hypochlorite in the

Gene name	*Description*	*S. sonnei start position*	*S. sonnei end position*	*S. flexneri start position*	*S. flexneri end position*
acrA	AcrAB-TolC pump gene[21]	477813	476620	420957	422150
acrB	AcrAB-TolC pump gene[21]	476597	473448	417785	420934
acrD	AcrD aminoglycosides resistance pump[22]	2689662	2692775	2575652	2578765
tolC	AcrAB-TolC pump gene[21]	3331624	3333105	3171250	3172731
emrB	EmrB efflux protein[1]	2973840	2975378	2786513	2788051
emrD	EmrD efflux protein[1]	3796932	3798122	3897623	3898675
ydhE	Transporter membrane protein[19]	1568520	1567147	1723184	1723866
marA	AcrAB expression[18]	1678452	1678069	1595773	1596156
SoxS	AcrAB expression[18]	4501826	4501503	4287490	4287813
acrR	Regulation of acrAB genes[23]	477955	478602	422292	422939
marR	egulation of acrAB genes[23]	1678906	1678472	1596553	1596176
cmr	Chloramphenicol resistance pump[24]	-	-	829096	830322

Table 1. Genes associated with membrane efflux pumps in *Shigella* species. Genes considered to have an association with the expression and regulation of efflux pumps were identified either through a manual search of current literature, or through keywords, such as 'Multidrug efflux' in reference strain annotations.

presence or absence of the efflux pump inhibitor CCCP. 1 mM CCCP stock solutions were prepared by dissolving solid 1.02g CCCP (Sigma-Aldrich, Dorset UK) in 5ml Dimethyl sulfoxide (DMSO) (Sigma-Aldrich, Dorset UK). Preparations of 2.5, 5 and 10 mg/l calcium hypochlorite solutions were added to 15ml centrifuge tubes, with 1 mg/l of CCCP solution added to tests with the inhibitor, to a total volume of 8ml for each test condition. Tubes were then inoculated with 2 ml of the tested bacterial strain and experiments run for a total of 30

minutes. Population counts were taken at 5, 15 and 30 minutes after neutralisation of free chlorine in solution with 100µl of 0.1M sodium thiosulphate ($Na_2S_2O_3$) (Sigma-Aldrich, Dorset UK).

Inactivation kinetics and statistical analysis

$CT_{99.9\%}$ values relating to the concentration of calcium hypochlorite required to inactivate bacterial populations by 99.9% were calculated for each *Shigella* strain, both in the absence and presence of the efflux pump inhibitor, as described in chapter 2. Paired *t tests* to compare $CT_{99.9\%}$ values between tests with and without CCCP were conducted in *R*[25].

Genomic analysis

Putative variation is genes reported previously to be involved in membrane pump activation (**Table 1)** will be identified through by analysing whole genome sequence data. Raw FASTQ paired read sequence data for each strain was obtained from the Wellcome Trust Sanger Institute online database[26]. Reads were aligned against either the *S. sonnei* Ss046 or *S. flexneri* str. 2a 301 reference strains using the Burrows-Wheeler Aligner (BWA)[27]. Single nucleotide polymorphisms (SNPs) were called in each sample against the reference strain using Samtools 'mpileup'[28] and low quality and low coverage variants (DP <10; GQ >30) were filtered with the vcfProcess function in *R v1.3.2* (**Supplementary materials).**

3.4 Results and discussion

Chlorine inactivation with efflux pump inhibition

To determine the role of efflux pumps in chlorine resistance in *Shigella* species, *in vitro* inactivation assays were conducted in solutions of varying calcium hypochlorite concentration, both in the presence and absence of the general efflux pump inhibitor CCCP. To quantify the biocidal activity of calcium hypochlorite on *S. sonnei* and *S. flexneri*, the average survival fraction of bacteria was calculated from population counts in triplicate after treatment with varying concentrations of the disinfectant. $CT_{99.9\%}$ values, a measure of amount of the disinfectant required to reduce the population by 99.9%, were obtained from the survival fractions, using the methods described in **chapter 2**, These values were calculated for each strain in the presence and absence of CCCP, and then compared to determine strain sensitivity to chlorination when efflux pumps are active and inhibited (**Table 2)**. Control experiments were also run with CCCP in the absence of calcium hypochlorite in the solution. These experiments showed that CCCP alone did not significantly alter the

Strain	*No CCCP*	*CCCP*
S. flexneri 1266	220.6171	135.7801
S. flexneri 262.78	166.4345	139.6551
S. flexneri 9.63	274.5846	126.5635
S. flexneri MS0052	138.1009	168.1002
S. flexneri DE0530	182.9997	153.4162
S. flexneri EG419	154.2753	158.1901
S. sonnei 54210	156.6956	144.8029
S. sonnei 88.83	176.6425	151.1814
S. sonnei 43.74	162.5593	162.5182
S. sonnei MS004	245.0319	204.4516
S. sonnei DE1208	215.2096	161.2225
S. sonnei EG0430	239.4178	220.2689

Table 3.2. $CT_{99.9\%}$ values for *S. sonnei* and *S. flexneri* strains tested in this study in the presence and absence of CCCP. Values calculated using linear regression model the rate coefficient of disinfection. (See **chapter 2** for more detailed description of the calculations used). All models fit the data significantly ($p < 0.05$).

survival of any *Shigella* strain (**Supplementary materials)**, thus any difference in chlorine resistance in a particular strain when in the presence of CCCP is likely due to the lack of efflux pump activity.

The results show that overall there was a significant difference in resistance to chlorination of *Shigella* when tested in the presence of the efflux inhibitor CCCP (Paired t-test; t = 2.5346, p-value = 0.02774). The average $CT_{99.9\%}$ value across all strains when efflux pump activity was uninhibited (absence of CCCP) was 194.38 (SD 43.26), and 160.51 (SD 27.22) with CCCP added. The effect of efflux pump inhibition was not consistent across all strains, with the greatest effect seen in European *S. flexneri* 9.63, with a decrease in $CT_{99.9\%}$ value from 274.58 to 126.56. In contrast, the average $CT_{99.9\%}$ value of the Vietnamese *S. flexneri* strain MS0052 increased when tested with CCCP, though when comparing the raw survival counts in replicates of this strain there was no significant difference between treatments with and without CCCP ($p = 0.765$). This suggests that in this strain there was no effect of efflux pump inhibition on chlorine resistance.

It appears that the addition of CCCP had the greatest effect on chlorine resistance in European *S. flexneri* and Vietnamese *S. sonnei,* with these groups also displaying the highest innate resistance with unihibated efflux pump activity (**Table 2 and chapter 2 results)**. It may be that increased efflux pump activity in these groups is causing a higher resistance to chlorine, with other systems involved in maintaining a baseline level of sensitivity. The inhibition of these pumps will return the levels of resistance to those conferred by these other processes.

Indeed, it has been reported that bacteria will have a variety of responses to chlorine, including changes to carbon metabolism and the activation of other stress responses, depending on the strength of the disinfectant[29]. It may be that raising the activity of efflux pumps is only one mechanism, chiefly employed by the organism to increase resistance at particularly high concentrations of chlorine whilst at lower, sub-lethal levels these other processes will encompass the initial response to weaker disinfection. The results of these experiments suggest that efflux pumps are most important in conferring resistance to high chlorine concentrations, and thus there may be a variety of mechanisms activated in *Shigella* in response to varying disinfectant strengths. Conducting gene expression profiling, such as DNA microarrays, to elucidate which genes are involved in this response will help to better understand the process of chlorine resistance.

Furthermore, CCCP is a generalised pump inhibitor that has been shown to constrain the activity of many common efflux pumps, in particular AcrAB-TolC, but will have variable efficacy against some other systems[30]. The activity of the AcrAB-TolC pump was the main focus of this chapter as it had been previously shown to be associated with bacterial chlorine resistance[15], though most microorganisms will contain a variety of additional generalised and substrate-specific transport systems. Again a measure of gene expression in these organisms with exposure to varying concentrations of chlorine will identify the extent that certain pumps are involved in chlorine resistance.

Fluoroquinolone resistance and chlorine disinfection

Fluoroquinolone resistance could be confidently predicted in the tested samples through mutations in the DNA gyrase gene *gyrA,* an essential component of the DNA supercoiling process[11,19], and phenotypic susceptibility data (**Chapter 2, Table 1)**. Key SNPs in codon 83 and 87 of the gene have been shown previously to confer a greater resistance to fluoroquinolones, with these mutations identified in *Shigella* populations within Vietnam[31,32] (**Chapter 4)**.

Of the 12 strains tested here, only two carried one of these key mutations at the *gyrA* locus, inferring resistance to fluoroquinolones, thus it proved difficult to accurately predict the effect of this trait on chlorine resistance. Strains *S. flexneri* EG419 and *S. sonnei* EG0430 from Vietnam harboured the mutations conferring drug resistance and did not appear to have a higher average level of resistance to chlorine disinfection than susceptible strains of same group (species and location).

Interestingly, the European *S. flexneri* strain 9.63 had the greatest difference in chlorine resistance when efflux pump activity was inhibited ($CT_{99.9\%}$ value 274.58 to 126.56), and also the highest individual resistance of all strains. Phenotypic drug susceptibility tests showed that this strain resistance to the highest number of antibiotics (**chapter 2, Table 1)** and thus is may be that the additive effect of multiple drug resistances is increasing efflux pump activity, which in turn will increase chlorine resistance. This theory is consistent with previous work describing the synergistic effects of multiple drug resistance on efflux pumps[33], though owing to the low number of strains tested here, further experiments would be required to assess the validity of this hypothesis.

Genomic variation in efflux pump genes

The analysis of whole genome sequences of the *S. sonnei* and *S. flexneri* strains used in this study identified variation in genes that had been previously associated with encoding and regulating efflux pumps, either in *Shigella* species or closely related Enterobacteriaceae, such as *E. coli* **(Table 1).** There were six non-synonymous single nucleotide polymorphisms (SNPs) in five known pump associated genes in *S. sonnei,* though none of these mutations appeared to be involved in an increased resistance to chlorine or a greater change in resistance with the inactivation of general efflux pumps. Additionally, no SNPs in pump genes were found only in *S. sonnei* EG0430, known to be resistant to fluoroquinolones.

The sequence analysis identified 39 non-synonymous mutations (55 total mutations) across *S. flexneri* isolates in 10 of the 12 selected pump genes. The increased number of mutations in *S. flexneri* compared to *S. sonnei* is likely due to the increased genetic diversity across *S. flexneri* populations (**Chapter 4).** There was an excessive of non-synonymous mutations in *acrB, cmr* and *acrD,* though the majority of these SNPs were only found in a single isolate and so it was difficult to determine the effect on chlorine sensitivity. The highest levels of resistance to chlorine were found in *S. flexneri* were in European strains 1266 and 9.63, and there were no mutations found only in these strains. The ratio of non-synonymous to synonymous mutations in these genes (39/16) does suggest that there may be selection driving variation in these regions, though the non-specificity of these pump systems and the

lack of associations with decreased susceptibility to chlorine makes it complicated to suggest putative variants that can be contributing to resistance in *Shigella.*

3.5 Conclusions

Efflux pumps found in the cellular envelope of a variety of bacterial taxa have been previously shown to be associated with increased resistance to a variety of toxic substrates, including antibiotics. Resistance to chlorine has also been proposed to be influenced by increased activity in these transport proteins. Results of inactivation assays in the presence of a generalised efflux pump inhibitor, CCCP, suggest that these proteins are involved in chlorine sensitivity in *Shigella*, with an increased efficacy of the disinfectant against strains that were previously resistant to higher concentrations of chlorine when the activity of these pumps is constrained.

The effects of pump inhibition were not consistent across all isolates tested and appeared to have the most impact on strains that were inherently more resistant to chlorine. This indicates that other stress induced processes will be activated with exposure to lower levels of chlorine. Additionally, these experiments were unable to resolutely find associations with chlorine sensitivity and resistance to fluoroquinolones and, though some variation in efflux pump genes was identified, this did not appear to correlate with differential chlorine resistance. Given the association between efflux pump activity and chlorine susceptibility demonstrated here in *Shigella,* the implementation of gene expression profiling techniques to identify genes that may be involved in the bacterial response to chlorine exposure would help to better understand the evolution of disinfectant resistance in these organisms.

References

1. Webber, M. A. & Piddock, L. J. V. The importance of efflux pumps in bacterial antibiotic resistance. 9–11 (2003). doi:10.1093/jac/dkg050

2. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13,** 42–51 (2015).

3. Wang, W. *et al.* High-level tetracycline resistance mediated by efflux pumps Tet(A) and Tet(A)-1 with two start codons. *J. Med. Microbiol.* **63,** 1454–1459 (2014).

4. Louw, G. E. *et al.* Europe PMC Funders Group Rifampicin Reduces Susceptibility to Ofloxacin in Rifampicin- resistant Mycobacterium tuberculosis through Efflux. **184,** 269–276 (2013).

5. Hirakawa, H. *et al.* AcrS/EnvR represses expression of the acrAB multidrug efflux genes in Escherichia coli. *J. Bacteriol.* **190,** 6276–6279 (2008).

6. Ruiz, J. Mechanisms of resistance to quinolones: target alterations, decreased accumulation and DNA gyrase protection. *J. Antimicrob. Chemother.* **51,** 1109–17 (2003).

7. Delmar, J. A., Su, C.-C. & Yu, E. W. Heavy metal transport by the CusCFBA efflux system. *Protein Sci.* **24,** 1720–1736 (2015).

8. Piddock, L. J. V. Multidrug-resistance efflux pumps ? not just for resistance. *Nat Rev Micro* **4,** 629–636 (2006).

9. Sun, J., Deng, Z. & Yan, A. Biochemical and Biophysical Research Communications Bacterial multidrug efflux pumps : Mechanisms , physiology and pharmacological exploitations. *Biochem. Biophys. Res. Commun.* **453,** 254–267 (2014).

10. Paulsen, I. T., Brown, M. H. & Skurray, R. A. Proton-dependent multidrug efflux systems. *Microbiol. Rev.* **60,** 575–608 (1996).

11. Marcusson, L. L., Frimodt-Møller, N. & Hughes, D. Interplay in the selection of fluoroquinolone resistance and bacterial fitness. *PLoS Pathog.* **5,** e1000541 (2009).

12. Pos, K. M. Drug transport mechanism of the AcrB efflux pump. *Biochim. Biophys. Acta* **1794,** 782–93 (2009).

13. Yuan, Q. Bin, Guo, M. T. & Yang, J. Fate of antibiotic resistant bacteria and genes during wastewater chlorination: Implication for antibiotic resistance control. *PLoS One* **10,** 1–11 (2015).

14. Prasad Karumathil, D., Yin, H. B., Kollanoor-Johny, A. & Venkitanarayanan, K. Effect of chlorine exposure on the survival and antibiotic gene expression of multidrug resistant Acinetobacter baumannii in water. *Int. J. Environ. Res. Public Health* **11,** 1844–1854 (2014).

15. Potenski, C. J., Gandhi, M. & Matthews, K. R. Exposure of Salmonella Enteritidis to chlorine or food preservatives increases susceptibility to antibiotics. *FEMS Microbiol. Lett.* **220,** 181–186 (2003).

16. Versace, B. Z. and I. Inhibitors of Multidrug Resistant Efflux Systems in Bacteria. *Recent Patents on Anti-Infective Drug Discovery* **4,** 37–50 (2009).

17. Ardebili, A., Talebi, M., Azimi, L. & Rastegar Lari, A. Effect of efflux pump inhibitor

carbonyl cyanide 3-chlorophenylhydrazone on the minimum inhibitory concentration of ciprofloxacin in Acinetobacter baumannii clinical isolates. *Jundishapur J. Microbiol.* **7,** 3–7 (2014).

18. Webber, M. A. & Piddock, L. J. V. Absence of mutations in marRAB or soxRS in acrB-overexpressing fluoroquinolone-resistant clinical and veterinary isolates of Escherichia coli. *Antimicrob. Agents Chemother.* **45,** 1550–1552 (2001).

19. Hooper, D. C. Mechanisms of fluoroquinolone resistance. *Drug Resist. Updat.* **2,** 38–55 (1999).

20. Swick, M. C., Morgan-Linnell, S. K., Carlson, K. M. & Zechiedrich, L. Expression of multidrug efflux pump genes acrAB-tolC, mdfA, and norE in Escherichia coli clinical isolates as a function of fluoroquinolone and multidrug resistance. *Antimicrob. Agents Chemother.* **55,** 921–924 (2011).

21. Du, D. *et al.* Structure of the AcrAB-TolC multidrug efflux pump. *Nature* **509,** 512–515 (2014).

22. Elkins, C. a & Nikaido, H. Substrate Speci city of the RND-Type Multidrug Ef ux Pumps AcrB and AcrD of. *Society* **184,** 6490–6498 (2002).

23. Pourahmad Jaktaji, R. & Jazayeri, N. Expression of acrA and acrB genes in Esherichia coli mutants with or without marR or acrR mutations. *Iran. J. Basic Med. Sci.* **16,** 1254–1258 (2013).

24. Nilsen, I. W., Bakke, I., Vader, A. & Olsvik, Ø. Isolation of cmr , a Novel Escherichia coli Chloramphenicol Resistance Gene Encoding a Putative Efflux Pump. **178,** 3188–3193 (1996).

25. R Development Core Team, R. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **1,** 409 (2011).

26. Wellcome Trust Sanger Institute. No Title. (2016).

27. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

28. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

29. Gray, M. J., Wholey, W.-Y. & Jakob, U. Bacterial responses to reactive chlorine species. *Annu. Rev. Microbiol.* **67,** 141–60 (2013).

30. Mahamoud, A., Chevalier, J., Alibert-Franco, S., Kern, W. V. & Pagès, J. M. Antibiotic efflux pumps in Gram-negative bacteria: The inhibitor response strategy. *J. Antimicrob. Chemother.* **59,** 1223–1229 (2007).

31. Holt, K. E. *et al.* Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17522–7 (2013).

32. Thompson, C. N. *et al.* Clinical implications of reduced susceptibility to fluoroquinolones in paediatric Shigella sonnei and Shigella flexneri infections. 807–815 (2016). doi:10.1093/jac/dkv400

33. Bollenbach, T. Antimicrobial interactions: Mechanisms and implications for drug discovery and resistance evolution. *Curr. Opin. Microbiol.* **27,** 1–9 (2015).

Chapter 4

The characterisation of antibiotic resistance genes and compensatory mutations in *Shigella sonnei* and *Shigella flexneri* in Vietnam

4.1 Abstract

The acquisition of antimicrobial resistance in bacterial pathogens is of great concern to global human health. The fitness benefits of harbouring a resistance gene can allow a particular strain to spread rapidly in an area, with the beneficial gene also becoming more prevalent. There has been evidence, though, of some costs associated with mutations at chromosomal resistance loci through the loss or disruption of original gene functions. Recently there has been growing evidence of additional mutations in resistant strains that will restore fitness to high levels in the absence of the antibiotic, termed 'compensatory mutations'.

Shigella sonnei has been rapidly emerging as a major disease causing pathogen in many developing countries, replacing *S. flexneri* as the primary cause of dysentery in these regions. Several hypotheses have been suggested to explain this process, with the development of antibiotic resistance proposed as contributing factor in the spread of pathogens. This work aims to characterise key resistance mutations in Vietnamese populations of *S. sonnei* and *S. flexneri* by identifying single nucleotide polymorphisms (SNPs) that may be under selection. Variation in previously described resistance genes is also detected in order to determine the resistance profile of all *Shigella* strains. Finally, variants that are significantly associated with harbouring drug resistance are identified to search for putative compensatory mutations.

The results of this work show that drug resistance genes were identifiable in both *S. sonnei* and *S. flexneri,* with the latter species possessing a greater potential number of resistance conferring mutations but a lower proportion of resistant strains. From these variants, resistance to fluoroquinolones could be accurately predicted for each strain, with *S. sonnei* having a larger proportion of resistant isolates in the tested sample. Finally, no variants that were significantly more likely to be found in resistant or susceptible strains could be proposed as potential compensatory mutations.

4.2 Introduction

Antibiotic resistance in bacteria is among the greatest challenges facing global health, with estimates of around 2 million people infected by antibiotic resistant pathogens each year[1]. With the number of resistant bacteria increasing[2,3], there is a growing need to investigate the aetiology of strains harbouring these phenotypes in order to better influence effective strategies to control the spread of resistance, and to inform the development of new drugs. One such strategy to constrain the spread of resistance genes among populations is to limit the circulation of a particular antibiotic from a region for a period of time to lessen the selective pressure on bacteria to develop resistance. The acquisition and maintenance of chromosomal resistance mutations has been shown to have a deleterious effect on the fitness of an individual in the absence of the antibiotic, thus in an antibiotic-free environment susceptible individuals will possess a higher fitness[4]. It is theorised that by removing an antibiotic from circulation, resistance will be reversed in a population, and re-introduction at a later time will be effective in fighting the disease.

The recent identification of associated mutations in resistant bacteria that can compensate for these fitness costs eliminates the efficacy of this approach[5,6]. The specifics of compensatory mutations are detailed in Chapter 1 of this thesis. The benefit of these mechanisms that mitigate this reduction in fitness suggests that harbouring resistance and compensatory mutations together could, therefore, be a major factor influence on the potential for a particular species or strain of pathogen to successfully persist and spread in areas of increasing antimicrobial use. This study will focus on identifying putative antibiotic resistance and compensatory mutations in Vietnamese populations of *S. sonnei* and *S. flexneri,* and in particular on the potential for these processes to be driving the expansion of the *S. sonnei* in this region.

The impact of antibiotic resistance of the evolutionary history of *Shigella* has been explored previously, both within Vietnam and globally. Using whole genome sequences of 132 *S. sonnei* isolates collected between 1943 and 2010 from over 20 countries (including 19 from Vietnam), Holt *et al* reconstructed the phylogeny of the species using maximum-likelihood[7]. This analysis determined a most recent common ancestor (MRCA) of all isolates of less than 500 years ago and also sorted *S. sonnei* into four distinct lineages corresponding to different emergence times. The study examined the distribution of known antibiotic resistance mutations in the phylogeny and found that global dissemination is strongly linked with the development of multidrug resistance (MDR). The evidence for independent acquisition of the same resistance mutations among different clades in the tree, along with little evidence for positive selection in other genes, suggests strong selective pressure in these genes. In

lineage III, the globally distributed clade to which the majority of Vietnamese isolates belong, there is evidence for the plasmid-mediated acquisition of class 2 integrons (In*2*) immediately prior to international dissemination, as well as some cases of point mutations conferring quinolone resistance in the chromosomal *gyrA,* DNA gyrase, gene. This suggests that the first *S. sonnei* strains to reach Vietnam already had MDR to a number of antibiotics, which would have given them a selective advantage over other susceptible pathogens. Evidence suggests that, although antibiotics may not actually resolve infections caused by *S. sonnei,* treatments can prevent the shedding of the bacteria after the symptoms have resolved, preventing transmission between humans[8]. Thus resistance phenotypes may be important in sustaining the bacteria within host populations and facilitating its spread.

A further study from Holt *et al* sequenced the whole genomes of 244 isolates of *S. sonnei* from three regions in Vietnam between 1995 and 2010 and reconstructed their evolutionary history, including the 19 Vietnamese isolates from the previous study[9]. This analysis traced the MRCA of all *S. sonnei* in this study to a single MDR ancestor around 1982, suggesting either a novel introduction to the region or a historical bottleneck, with the greater support for the introduction hypothesis. The inclusion of isolates from three geographic areas in this study allowed for phylogeographic analysis to explore the spread of *S. sonnei* within the region. Isolates were collected from three provinces with the majority of dysentery cases in Vietnam, Ho Chi Minh City (HCMC), Hue and Khanh Hoa province (KH). It was found that the MRCA of the Hue and KH isolates was closely linked to the HCMC isolates and a large number of these strains were clonal, suggesting local recent clonal expansion into these areas from HCMC. There were also multiple KH and Hue isolates scattered throughout the phylogeny indicating separate events where *S. sonnei* has spread to these areas. These isolated events were only rarely successful in establishing new populations though (around 10% of establishments), pointing to potential differences in fitness and transmissibility between strains.

Closer inspection of the evolution dynamics of the HCMC isolates revealed four population bottlenecks which coincided with increases in the proportion of reported dysentery cases caused by *S. sonnei* and reports of increases in antibiotic resistance. Genetic data found that in these bottlenecks, a subset of accumulated mutations became fixed in the local population whilst others were lost. This indicates four incidences of selective sweeps, an elimination of variation around strongly selected mutations, where selection is driving competition between strains and the emergence of successful clonal populations. Single nucleotide polymorphism (SNP) analysis of plasmid mutations revealed four SNPs that became fixed in the first bottleneck and one in the second sweep, including the *ipgD* gene encoding an effector protein involved with cellular invasion[10].

Chromosomal SNPs were distributed evenly across the genome showing no general pattern of selection. Some mutations were found to have been independently acquired more recently among different populations, including point mutations in the DNA gyrase *gyrA* gene, associated with fluoroquinolone resistance, that became fixed during the third selective sweep. This more recent evolution of antibiotic resistance, along with evidence of further differentiation of from the accessory genome among isolates from the three geographic regions, suggests that these populations have continued to evolve since the introduction of their MRCA into Vietnam, which may have contributed to their ability to expand into new areas. With this pathogen becoming a growing concern for human health in the region, it is important to understand the mechanisms, including the accumulation of resistance genes, by which particular strains are able to successful spread and become established.

A recent study detailing the evolutionary history of *S. flexneri* on the other hand, did not find evidence of any significant global dissemination of antimicrobial resistant populations of the pathogen in spite of multiple occurrences of local acquisition of resistance genes[11]. It was reported that *S. flexneri* populations were likely to be remain localised for long periods of time without intercontinental spread, in contrast to the patterns common in *S. sonnei* expansions. Again, the evolutionary history of the species was reconstructed, finding that *S. flexneri* strains could be classified into seven, highly diverse lineages. It was proposed that the long term maintenance of diverse pathogen populations in smaller geographic regions would lessen the selective advantage of transient antimicrobial resistance gene acquisition in *S. flexneri*, and as a result there are a greater proportion of *S. flexneri* strains are less likely to carry these mutations or genomic elements conferring resistance. Thus, it may that the recently emerged *S. sonnei* populations will contain a greater number of resistant strains as there is a greater selective pressure on the organism to acquire resistance for successful persistence and spread of particular strains. With the increasing use of antibiotics in developing countries in recent history[12], the higher proportion of resistant *S. sonnei* strains be accelerating its spread over *S. flexneri.*

Additionally, it is possible that compensatory mutations may be increasing the fitness of *S. sonnei* strains in Vietnam. Experimental work carried out on the fitness effects of antibiotic resistance genes in *S. sonnei* from Vietnam showed that resistance to fluoroquinolones does not reduce the selective fitness of strains carrying resistant gene copies when competing with wild-type antibiotic-susceptible strains in mixed cultures (S. Baker 2013, pers comm., 14th May). This could suggest either that there are some epistatic interactions between genes or compensatory mutations that will be restoring the fitness of resistant bacteria. Whilst epistasis in genes is not being ruled out, it is likely that compensatory mutations are

contributing somewhat to this increase in fitness as this has been observed numerous occasions in other bacterial pathogens[13–16], as well as in theoretical models of gene and function losses[17]. If these compensatory mechanisms are characteristic of *S. sonnei* only, co-evolution with antibiotic resistance genes may contribute to the observed spread of *S. sonnei,* and associated decrease in *S. flexneri* infections, in developing regions.

The primary focus of this chapter is to gain a further insight into the genetic features that have supported the rapid emergence of *S. sonnei* and the replacement of *S. flexneri* as the primary cause of dysentery in Vietnam, discussed in detail in Chapter 1. Based upon previous findings on the global spread of each species, it appears that there may be key differences in the impact of resistance on their chances of successful establishment, with evidence that *S. sonnei* expansion will be in part driven by the acquisition of these genes. Potential resistance genes in the core chromosomal genome of *S. sonnei* and *S. flexneri* isolates will be identified from whole genome sequence data through a combination of techniques, with putative mutations involved in the compensation of fitness costs associated with resistance also targeted.

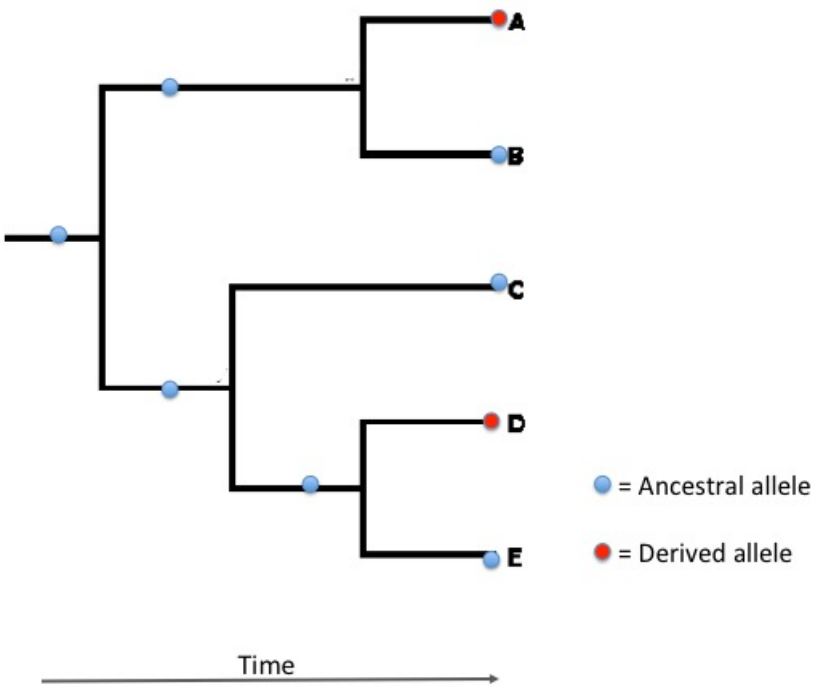


Figure 1. Diagrammatical representation of convergent evolution across a phylogenetic tree. The red allele is carried by individuals A and D, whilst all other individuals carry the blue copy. Reconstruction of ancestral state of the allele across the branches of the tree reveals that blue is the ancestral version of the gene, thus the red allele will have arisen independently through convergent evolution at tips A and D and is a homoplasy.

Genomic regions under selection will be identified by maximum parsimony analysis, identifying mutations that have arisen at multiple independent times across a population, and manual inspection of known, high-confidence resistance genes. Widely used techniques such as the ratio of non-synonymous to synonymous mutations (*dN/dS),* are not suitable for detecting recent positive selection in highly related populations such as *S. sonnei*[18]*.* Parsimony though can be used to identify homoplastic SNPs that have evolved through convergent evolution **(Figure 1)**[19], and this has been shown to be an effective tool for distinguishing areas of the genome that are under strong, recent selection in species of closely related pathogenic bacteria[19,20]. The genomes of *Shigella* species are known to be relatively stable[11], with little recombination and thus gene variants that have arisen in multiple, independent linages within a population will likely be driven by some selective pressure. This method can also detect SNPs non-genic regions of the genome, such as promotors, that may be under selection. The aim is to use these tools to detect regions of *S. sonnei* and *S. flexneri* chromosomal genomes under selection, and link any variation to traits that may be associated with antibiotic resistance within the organism. This allows for resistant and susceptible isolates for a particular antibiotic to be identified within each population, and putative compensatory mutations can be revealed by common gene variants found in a higher frequency in the resistant populations.

4.3 Methodology

Shigella isolate collection

High coverage whole genome sequences of 134 *Shigella flexneri* and 145 *Shigella sonnei* clinical isolates were obtained from the Wellcome Trust Sanger Institute online database[21]. Strains were collected from patients in three regions of Vietnam, Ho Chi Minh, Hue and Khanh Hoa between 1995-2010 as detailed in Holt *et al*[9]*.* All *S. sonnei* isolates were members of the global lineage III clade, and *S. flexneri* samples were predominately lineage 2a, the most common group within Vietnam. Further information including the date of isolation, location and individual accession numbers can be found in the **supplementary materials.**

Assembly and genomic analysis

Paired read FASTQ files were quality checked with adapter removal and read trimming performed with Trimmomatic[22]. Reads were then aligned against either *S. sonnei* Ss046 or *S. flexneri* 2a. 301 reference strains using BWA -mem and -sampe programs (v.0.7.12)[23],

with the median coverage across *S. sonnei* and *S. flexneri* alignments 150x and 125x respectively. Aligned sequences were converted to BAM format with Samtools (v1.3.1)[24] and multi-sample variant calling implemented with the 'mpileup' command in the same program.

SNPs were discarded from problematic repeat regions and filtered for low quality (QC <30) and low coverage (DP <10) variants using vcfProcess, a *R* function developed for post-processing variant caller outputs (**Supplementary materials).** To maximise the genetic variation identifiable in samples, mixed (heterozygous) calls were assigned as the reference or alternative allele if supported by >80% of reads at a site, otherwise these were marked as a gap. SNPs falling <10bp apart were excluded from the phylogenetic input but considered when identifying genes under selection. Sequences of high confidence SNPs were then concatenated for phylogenetic and selection analyses.

Phylogenetic analysis

High confidence SNPs were used to reconstruct the maximum-likelihood phylogeny for *S. sonnei* and *S. flexneri* isolates using RaxML[25], using the GTRGAMMA model of base substitution and site heterogeneity. 1,000 bootstrap replicates were run to calculate support for the final tree topology of each species. Trees were visualised with selected SNPs of interest mapped against tips using a bespoke script in the R package 'ape'[26].

Identification of genes under selection

Homoplastic SNPs that have arisen multiple times over each species' phylogeny were identified using the 'phangorn' package in *R*[27]. The ancestral state of sites was reconstructed at each node by 'Fitch's' parsimony algorithm and the number of mutational changes forward through time that have taken place across the tree to result in the distribution of derived alleles at tips is calculated for each SNP. A figure of >1 at any site indicates at least two independent mutations across the tree, evidence of convergent evolution. The resulting homoplastic SNPs were mapped to gene regions using the reference strain annotation, and the likelihood of conferring antibiotic resistance or fitness compensation considered by determining gene function using the PANTHER database[28].

Resistance genes were also identified through a search of current literature and were manually examined for genetic variation within *Shigella* genomes. Finally, *S. sonnei* and *S. flexneri* strains were tested for likely resistance to fluoroquinolones *in silico* using ARIBA[29]. With these groupings, allele frequency differences between resistant/susceptible isolates were used SNPs that were found in a significantly higher frequency in isolates harbouring a particular resistance conferring mutation and thus may be a candidate compensatory

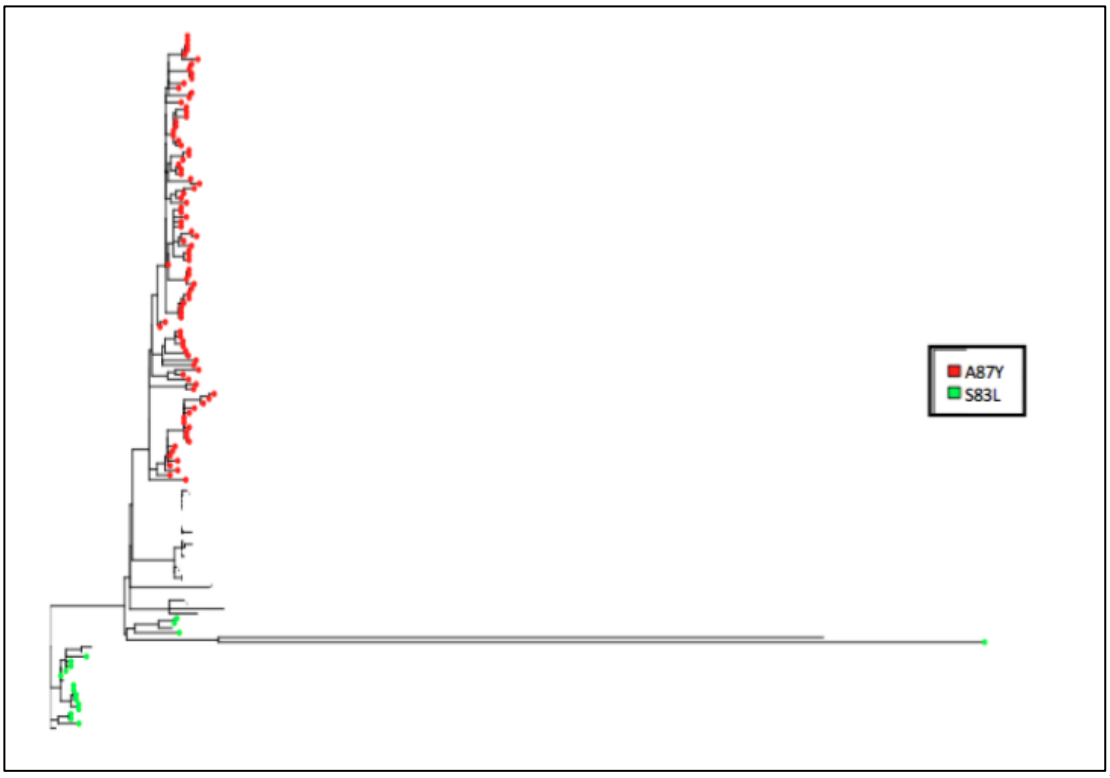
mutation, through Discriminate Analysis of Principal Components (DAPC), implemented in the '*adegenet*' package in *R. v1.3.2*[30]*.* This will search for SNPs that contribute most to the pre-defined resistant/susceptible groups, while controlling for any population structuring within these groups that may give rise to false-positive associations due to the underlying evolutionary relationships within the groups.

4.4 Results and Discussion

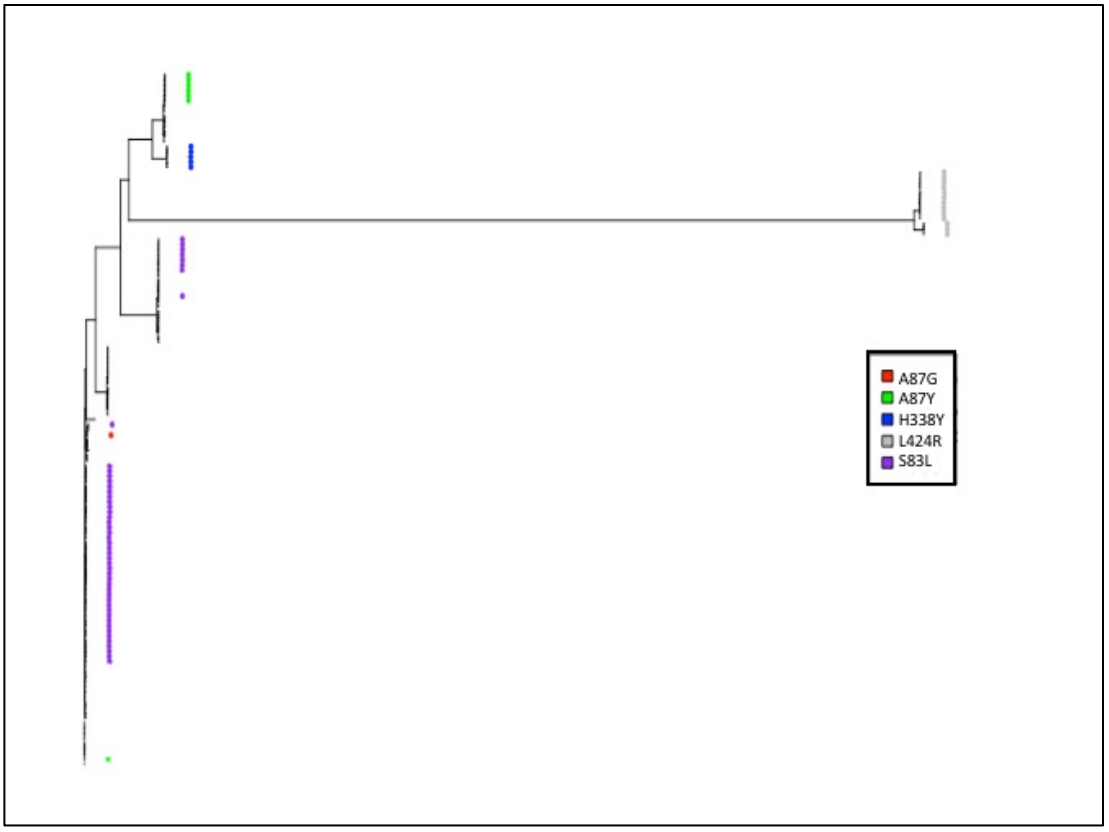
Population structure and genetic variation

Species-level population structure and evolutionary relationships in Vietnamese *S. sonnei* and *S. flexneri* were determined by conducting phylogenetic analysis on 145 *S. sonnei* and 134 *S. flexneri* whole genome sequences. Sequence reads were aligned to either the *S. sonnei* Ss046 or *S. flexneri* 2a str301 reference genome, and genomic variants detected between isolates. The number of SNPs found within each species is indicative of the disparity in genetic diversity between populations of *S. sonnei* and *S. flexneri,* with 1,050 SNPs identified in *S. sonnei* and 39,571 in *S. flexneri*. Of the variants found in *S. sonnei,* 412 were found in single isolates (39.24%), while only 0.026% (1,040) of SNPs in *S. flexneri* were present in only one sample. The average pairwise SNP distance between individuals was 37.42 in *S. sonnei* isolates and 8,604.13 between *S. flexneri*. The difference in the levels of variation between individuals of each species is in line with previous evidence of more recent, clonal evolution of *S. sonnei* in Vietnam in contrast to *S. flexneri* populations, which can be characterised by distinct, deeply-rooted lineages, with relatively high diversity both between and within these groups.

A robust maximum-likelihood phylogeny of high quality SNPs across each species was built to infer evolutionary relationships between isolates (**Figures 2 & 3).** The resulting tree topologies also reflect the differing population structure between *Shigella* species. *S. sonnei* isolates largely form a single monophyletic clade with low diversity, with the individuals falling outside of this group representing the reference strain and one individual isolated in 1995. In contrast, *S. flexneri* forms at least five distinct clades, likely distinguishing lineage differences, that are highly divergent. There are 13 *S. flexneri* isolates that form a very distant branch of the tree from all other strains, suggesting a deep-rooted evolutionary divergence. These samples were collected across all three sites in Vietnam between 2003-2010 and so will not likely be due to a single outbreak of closely related infections. This illustrates that the major source of genomic diversity in *S. flexneri* will be due to lineage specific mutations, which supports previous evidence of high diversity in this species[11].



A



B

Figure 2. Unrooted maximum-likelihood phylogeny produced by RaxML of (A) *Shigella sonnei* and (B) *Shigella flexneri* strains used in this study. All major nodes have >95% bootstrapping support. Mutations in the DNA gyrase gene, *gyrA*, which have been shown to confer resistance to fluoroquinolones are plotted on the tree tips and the gene coordinates shown in the legend.

Identification of homoplastic sites

Sites under positive selection were identified through reconstruction of the most parsimonious ancestral states of SNPs at each node of the maximum-likelihood tree, with homoplasies defined as occurrences of multiple mutations at a site given the sequence at the tips. The number of homoplastic sites in *S. flexneri* and *S. sonnei* populations varied markedly. There were six SNPs identified in the *S. sonnei* population that appear to have evolved on multiple, independent occasions. Of these sites, three are found in protein coding regions, with one non-synonymous mutation, S83L, in *gyrA* identified in 23/145 isolates and known to confer fluoroquinolone resistance, described previously in Vietnamese *S. sonnei* populations[9,31] (**Figure 2A).**

The other homoplastic SNPs in protein coding regions of *S. sonnei* were identified as a synonymous mutation in 18/145 isolates found in the heat shock protein *yceA,* and a double synonymous mutation (A>G in 144/145 and A>T 1/145 isolates) in *metR,* a regulatory protein of the pathway involved in methionine metabolism[32]. It is improbable that these mutations are involved in antimicrobial resistance. Similarly, the remaining three homoplastic SNPs were found in the same non-coding region of the genome, though due to their distance from the nearest gene and close proximity it is likely they are a result of a previously unidentified mobile element.

In contrast, there were 413 homoplastic SNPs identified across *S. flexneri* isolates, with 300 falling in coding regions. Of the SNPs in coding regions, 120 were non-synonymous mutations, altering the amino acid sequence of the gene. Mutations were found again in *gyrA* (A87Y, 7/134 isolates; S83L, 47/134 isolates), conferring fluoroquinolone resistance (**Figure 2B),** with these mutations previously identified in Vietnamese *S. flexneri* populations[33]. 55 homoplastic genes were assigned molecular function GO terms using the PANTHER classification tool[28]. The majority of these genes were termed catalytic activity (GO:0003824) (26/55 genes) and eight were involved in transporter activity (GO:0005215), though there is no previous evidence of antimicrobial activity associated with any of these genes.

Six non-synonymous SNPs were nonsense mutations, resulting in a premature ‘STOP’ codon in a gene that will arrest the amino acid sequence and cause pseudogenisation. One such gene containing a nonsense mutation was in the *mdaA* gene (9/134 isolates), which has been shown in *E. coli* systems to be associated with increased resistance to tumoricidal drugs when over-expressed, including DMP 840[34]. In addition, *mdaA* has been shown to confer resistance to nitrofurantoin, a second line antibiotic used to treat bladder infections[35]. Interestingly, the gene is controlled by the MarA regulon, shown to also regulate the activity

of the multidrug efflux pump AcrAB-TolC[35]. It is possible that the loss of function in the *mdaA* gene may be associated in some way either with resistance to one of these compounds, or having an epistatic relationship with the genes involved in efflux pump regulation, though more work would be required to investigate any consequence of this function loss.

Variation at known resistance loci

Genes that have been previously shown to be associated with antibiotic resistance, either in *Shigella* species or other Gram-negative bacteria, were also manually inspected for variants in *S. sonnei* and *S. flexneri* genomes. Only non-synonymous SNPs were considered in this analysis as it is known that these mutations will change the amino acid sequence of the coding region. 19 genes were chosen for further analysis based upon evidence of an association with either specific drug resistance, such as mecillinam susceptibility in *aroK,* or broad-spectrum effects, for example the genes encoding the AcrAB-tolC drug efflux pump.

There were six non-synonymous SNPs found in known resistance conferring genes in *S. sonnei* strains. Along with the homoplastic non-synonymous mutation discovered in the DNA gyrase gene through parsimony analysis, a further *gyrA* mutation, A87Y, was identified in 93/145 *S. sonnei* isolates. In addition, looking at heterozygous variant calls in the *gyrA* gene, a further isolate was identified that appeared to constitute a mixed infection of resistant and susceptible bacteria, with the A87Y mutation present in around 60% of the raw sequence reads. Phenotypic susceptibility information on this sample showed resistance to fluoroquinolones (**Supplementary materials)** and so this strain was characterised as resistant for the DAPC analysis presented later in this chapter.

In addition, there were two mutations in another gene known to be associated with higher levels of resistance to fluoroquinolones, *parC*. One variant was identified in 144/145 isolates, with a single sample harbouring an additional mutation in the gene. The remaining strain showed a heterozygous call at this locus, again suggesting a mixed infection. The ubiquity of this mutation across the population containing both fluoroquinolone resistant and susceptible strains suggests that this mutation in isolation will not confer resistance to this class of antibiotics, but there may be some epistatic or additive effect with the mutation when in combination with further resistance genes in *gyrA.* Research has shown the efficacy of fluoroquinolones can vary against *E. coli* carrying only variants in the *parC* gene with no *gyrA* mutations[36], with particular SNPs only having a weak effect on the MIC of these antibiotics if situated outside of the resistance-determining region.

There were 47 non-synonymous SNPs identified in previously described resistance genes in Vietnamese *S. flexneri* isolates, including the same *gyrA* and *parC* loci as were detected

Gene Name	***Antimicrobial Activity***	***S. sonnei start position***	***S. sonnei end position***	***No. of S. sonnei SNPs (no. of isolates)***	***S. flexneri start position***	***S. flexneri end position***	***No. of S. flexneri SNPs (no. of isolates)***
aadA	Aminoglycoside resistance[41]	4110458	4111246	0 (0)	-	-	0 (0)
acrA	AcrAB-TolC multidrug efflux pump[40]	477813	476620	0 (0)	420957	422150	**6 (68)**
acrB	AcrAB-TolC multidrug efflux pump[40]	476597	473448	**1 (1)**	417785	420934	**7 (134)**
ampC	*Resistance to beta-lactam antibiotics*[31]	*4604316*	*4605449*	*0 (0)*	*4482197*	*4483330*	**5 (53)**
ampD	Regulation of *ampC*[32]	127961	128512	0 (0)	116997	117548	**1 (13)**
ampE	Regulates of *ampC*[32]	128509	129363	0 (0)	117545	118399	**3 (14)**
aroK	Mecillinam resistance[33]	3681190	3681711	0 (0)	3487756	3488478	**2 (21)**
gyrA	Fluoroquinolone resistance[34]	2408970	2411597	**2 (116)**	2352232	2354859	**5 (73)**
gyrB	Fluoroquinolone resistance[35]	3818616	3821030	0 (0)	3872637	3875051	**1 (13)**
marA	Multiple antibiotic resistance[37]	1678069	1678458	0 (0)	1596156	1595773	0 (0)
marB	Multiple antibiotic resistance[38]	1677819	1678037	0 (0)	1595523	1595741	**1 (65)**
marR	Overexpression of *marA*[37]	1678472	1678849	0 (0)	1596176	1596553	**3 (133)**
parC	Fluoroquinolone resistance[36]	3320041	3322299	**2 (144)**	3159155	3161413	**5 (107)**
parE	Fluoroquinolone resistance[35]	3327013	3328905	0 (0)	3166639	3168531	**3 (53)**
rpoA	Compensation of *rpoB*[16]	3612840	3613829	0 (0)	3424423	3425412	0 (0)
rpoB	Rifampicin resistance[16]	4400376	4404404	**1 (1)**	4200998	4205026	**2 (21)**
rpoC	Compensation of *rpoB*[13]	4404481	4408704	0 (0)	4205103	4209326	0 (0)
soxS	AcrAB-TolC multidrug efflux pump[39]	4501503	4501826	0 (0)	4287490	4287813	**1 (13)**
tolC	AcrAB-TolC multidrug efflux pump[40]	3331624	3333105	0 (0)	3171250	3172731	**4 (53)**

Table 1. Selected chromosomal genes with evidence for an association with antibiotic resistance in *S. sonnei* and *S. flexneri.* The number of non-synonymous SNPs identified within each gene is shown, along with the number of isolates of each species that harbour at least one derived allele.

in the *S. sonnei* population. In addition, there were three other SNPs found in DNA gyrase in *S. flexneri*, with mutations at A87G, H338Y and LL424R. The A87G variant has been previously associated with increased fluoroquinolone resistance in Vietnamese *Shigella*[9,37], though these other SNPs appear to be novel variants in this population. The extent to which they will increase resistance to fluoroquinolones would need to be investigated through drug susceptibility testing, though there does not appear to be support for resistance through mutations outside of the short, resistance-determining region that includes codons 83 and 87 in this gene[38].

Similarly, the majority of the mutations in *parC* were not found at previously described resistance-determining loci, though one mutation, G80A, in 21/134 isolates had been shown to increase the MIC of fluoroquinolones to *E. coli*[38]. Based on previous characterisation of this G80A *parC* variant, strains harbouring this mutation are proposed to have some level of resistance to fluoroquinolones and will be considered as such when grouping strains as resistant or susceptible for the identification of putative compensatory mutations. All isolates that harboured this variant also had a mutation in *rpoB,* a common rifampicin resistance gene[16]. Mapping these mutations on to the phylogenetic tree showed that they were found in one monophyletic clade, as such it is not possible to determine if they are functionally connected, though it is likely they have evolved through independent events. This clade comprised the oldest *S. flexneri* isolates considered in this analysis from between 1995-1996, and interestingly this same gene was found in a single *S. sonnei* strain, again dating from the mid 1990's. This suggests that at this time there was a circulating population of rifampicin-resistant *Shigella*, though this resistance did not continue in more recent populations.

A number of SNPs were also identified in genes associated with the AcrAB-TolC efflux pump, a system that has been shown to influence antibiotic resistance in Gram-negative bacteria. The results from **chapter 3** of this thesis show that these pumps can also have an effect on chlorine resistance, though it was found that Vietnamese *S. flexneri* were more susceptible to chlorination than *S. sonnei,* and so it is unclear to the extent that mutations in the pump genes *acrA* and *acrB* will influence antibiotic resistance.

Finally, multiple non-synonymous mutations were identified in a repressor gene of the multiple antibiotic resistance (mar) operon, *marR,* with all but one strain harbouring at least a single SNP in this coding region. This gene encodes a group of regulatory factors that are modulated by environmental stress signals, such as exposure to antibiotics[39]. They have been associated with the transcription of virulence factors in pathogens and the regulation of drug efflux pumps[40]. Closer inspection reveals that there are multiple non-synonymous

variants yet no synonymous mutations in this gene, suggesting that there may be ongoing selection acting at this locus. *S. flexneri* infections are often typified by more a severe disease than *S. sonnei*[41] and it may be that mutations in this gene are contributing to increased pathogenicity in this species.

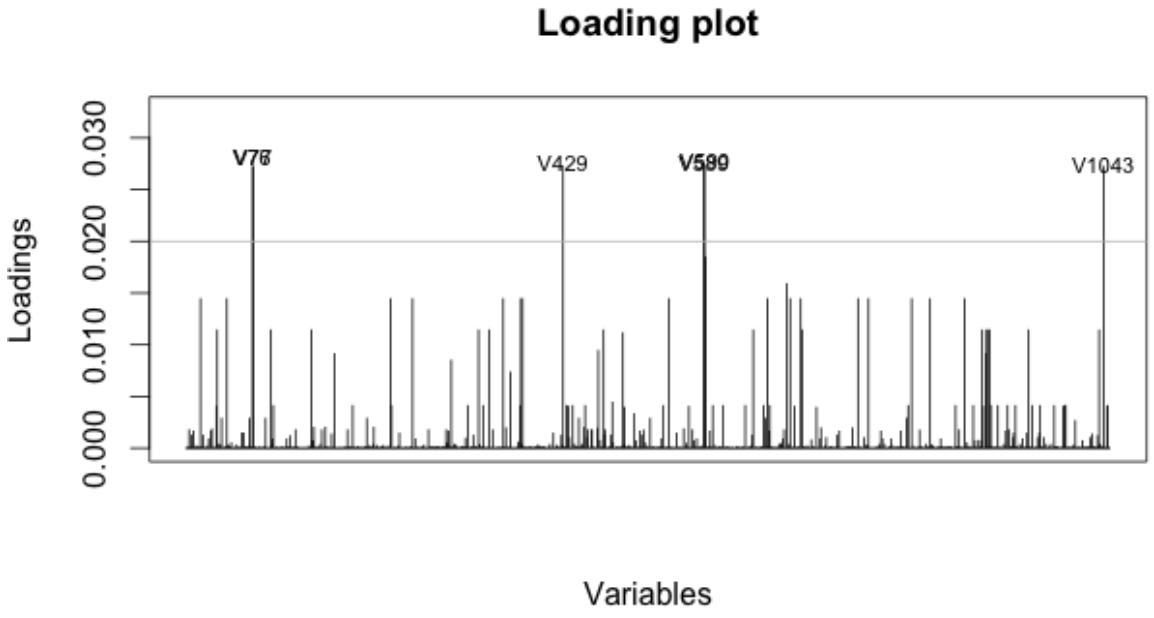
The detection of allelic variation in key resistance genes shows that there are differences in the diversity of these regions between *Shigella* species. Comparison of these genes between species is problematic, though, as variation will only be identified if there are intra-species nucleotide differences within the population, or variation between the tested isolates and the reference strain, thus fixed variants, irrespective of their selective benefits, will not be detected. To this end, it is difficult to assess all fitness differences informed by genomic variation between these species using this process in isolation, though it has been suitable to detect further informative resistance conferring variants in Vietnamese *Shigella* populations.

Compensatory mutations in fluoroquinolone resistance

The analysis of SNPs in whole genome sequences of *S. sonnei* and *S. flexneri* isolates from Vietnam identified mutations in resistance genes through the detection of homoplastic variants that have likely evolved due to selection, and by manually inspecting known resistance genes. Based on the characterisation of resistance genes in previous studies, and phenotypic susceptibility data in *S. sonnei*, I can confidently consider that variants identified here at codons 83 and 87 of *gyrA* in *S. sonnei* and *S. flexneri*, and codon 80 of *parC* in *S. flexneri* are likely to predict bacterial resistance to fluoroquinolones. Classification of strains is also confirmed by *in silico* testing of resistance using the ARIBA tool[29] (**Supplementary materials**), with 73/134 *S. flexneri* and 116/145 *S. sonnei* resistant strains.

Strains were separated into groups of either resistant or susceptible individuals within each species to search for evidence of compensatory mutations, as well as other associated variants. To identify mutations that are found in significantly higher frequencies in either susceptible or resistant strains, a Discriminant Analysis of Principal Components (DAPC) analysis was carried out comparing allele frequencies of all SNPs in these groups within each species (**Figure 4)**. Discriminating SNPs between fluoroquinolone-resistant and –susceptible *S. sonnei* were identified that revealed five mutations in coding regions and one in a non-coding region that were significantly more likely to be associated with one of the groups (**Table 2A)**. Intuitively, the most common *gyrA* variant in this population, A87Y, was detected through this process. Two further non-synonymous mutations were identified, though neither were candidates for compensatory mutations based on their annotated functions (*proY* encodes a proline transporter protein and *yfaA* is uncharacterised).

(A)



(B)

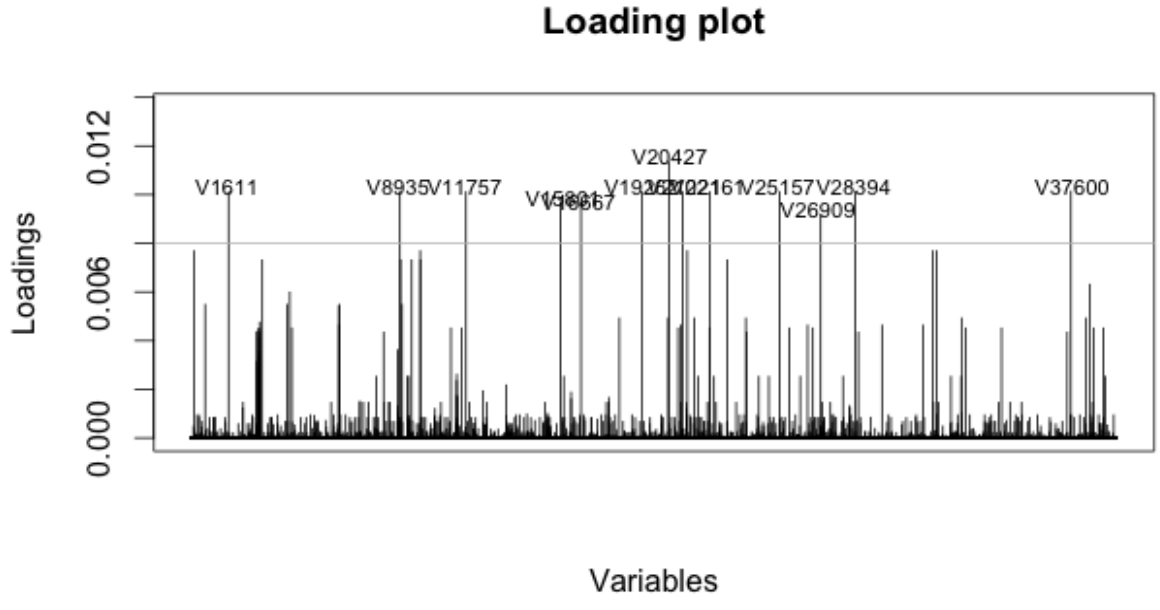


Figure 4. Discriminant Analysis of Principal Components identifying discriminating variables (SNPs) for each group (resistant/susceptible to fluoroquinolones). (A) DAPC plots for *S. sonnei* based on the top highest PCs contributing to group discrimination and a contribution threshold of 0.02. **(B)** DAPC plots for *S. flexneri* based on the highest 20 PCs contributing to group discrimination and a contribution threshold of 0.008. The SNP ID of discriminating SNPs over the threshold are shown.

(A)

SNP position	*Gene name*	*dN or dS*
406723	***proY***	non-synonymous
416501	***tsx***	synonymous
1624522	***ydgA***	synonymous
2407844	***yfaA***	non-synonymous
2411339	***gyrA***	non-synonymous

(B)

SNP position	*Gene name*	*dN or dS*
164493	***fhuB***	synonymous
1357086	***aldH***	non-synonymous
2424196	***pta***	synonymous
1034324	***ymcA***	non-synonymous
1866734	***zntB***	non-synonymous
1997465	***fliA***	non-synonymous
2354612	***gyrA***	non-synonymous
2540695	***cysP***	non-synonymous
2877071	***ygcX***	non-synonymous
3265586	***yhaU***	non-synonymous

Table 2. Gene name and position of SNPs that were found to be discriminating variables in (A) *S. sonnei* and (B) *S. flexneri* DAPC analysis.

The results of the discriminant analysis in *S. flexneri* resistant and susceptible groups identified three SNPs in *S. flexneri* in non-coding regions that were situated a relatively large distance from the nearest gene, so were unlikely to have any functional effect on each group. Of the remaining ten SNPs, two were synonymous mutations and eight non-synonymous, indicating potential functional differences in these genes between resistant and susceptible groups. Again as was expected, a mutation in the *gyrA* gene was identified in this analysis, the S83L variant, which is the most common variant in the group. Of the remaining non-synonymous mutations associated with each group, the most notable were three genes that encoded cellular transport proteins, *cysP, zntB* and *yhaU.* These genes were all linked to very specific systems for transporting substrates across the cell membrane[42,43] and so it is again unlikely that these could be considered as putative compensatory mutations.

As demonstrated by the detection of only one *gyrA* variant in each species with this discriminant analysis, only single sites are considered when looking for associations and

thus will not identify all the genomic variation contained in a population. If there are multiple variant loci in a gene that can imply similar functional differences, then these may not be picked up, with each SNP is considered a single variable. These methods, though, are robust for distinguishing sites under selection, and thus it is unlikely that there are any compensatory mutations associated with fluoroquinolone resistant in *Shigella* populations in Vietnam.

4.5 Conclusions

The aim of this study was to identify likely antibiotic resistance mutations from whole genome sequences of *S. sonnei* and *S. flexneri* strains in Vietnam by detecting sites under strong selection and variants in known resistance associated genes to determine possible differences in drug resistance profiles between species. In addition, mutations in resistant strains that would mitigate any fitness costs associated with antibiotics were targeted through discriminate analysis of variants in resistant and susceptible groups in each species.

Fluoroquinolone resistance could be confidently predicted in both species through well described mutations in the DNA gyrase gene, *gyrA,* an essential gene involved in transcription and DNA supercoiling[15]*,* as well as signals of strong selection at this gene in both species through the detection of homoplastic sites. These classifications were also supported by *in silico* testing for resistance. Non-synonymous SNPs in codons 83 and 87 of this gene have been shown repeatedly to confer an increased resistance to a range of fluoroquinolones[31,44], with these mutations also previously identified in Vietnamese *Shigella* populations[9,33]. The S83L variant in *gyrA* was found as an example of convergent evolution in both species, with this specific mutation found in the highest number of fluoroquinolone resistant *S. flexneri* strains. This point mutation has been found often to be the first novel *gyrA* mutation to spread in a population as it confers high levels of resistance with low functional changes to the DNA supercoiling[45]. A greater proportion of *S. sonnei* strains tested here (116/135) harboured resistance to fluoroquinolones than the *S. flexneri* population. This may be simply due to sampling bias or it may indicate some differences in the circulating resistance to this class of antimicrobials in Vietnamese *Shigella* species.

There were no further sites that appeared to be under strong selection in either species that could confidently be connected to antimicrobial resistance, though manual inspection of previously described resistance genes did identify additional variants that may be linked to resistance. Most interesting were SNPs in *parC,* a topoisomerase gene involved in DNA coiling[38,46], detected in both *Shigella* species and known to also play a role in fluoroquinolone

resistance[44,45]. One mutation in *S. sonnei* was at fixation within the population, yet phenotypic susceptibility testing showed that a number of isolates did not have resistance to nalidixic acid, a quinolone (**Supplementary materials**). This suggests that this mutation alone does not confer resistance, with further inspection revealing that the position of this SNP is outside the proposed resistance determining region of the gene[38]. Double mutants carrying both a *parC* and *gyrA* mutation in *Salmonella enterica* have been shown to have a fitness advantage over single mutants and wild type strains[47], thus whilst *parC* may not have evolved as a compensatory mutation, epistatic interactions between the genes may be mitigating fitness costs.

Similarly, five non-synonymous mutations were found in the *parC* gene in *S. flexneri*, though four of these again fell outside of this resistance region. One variant though had been previously found to confer a level of resistance to quinolones even in the absence of *gyrA* mutations and thus was proposed to determine resistance in the strains carrying it here.

S. flexneri appeared to have a greater range of potential resistance determining mutations across all isolates than *S. sonnei*, though a greater proportion of *S. sonnei* strains harboured at least one resistance gene. This supports the theory that the relatively long persistence of *S. flexneri* populations in an area lessens the selective advantage of harbouring transient resistance mutations. Thus these mutations are less readily acquired in a population but will remain as standing variation in the species even when the selective advantage is reduced[11]. The successful expansion of *S. sonnei* lineages, on the other hand, will be driven by acquisition of resistance mutations, and with the increased use of antibiotics in developing regions[12], the propensity for resistance to spread through *S. sonnei* populations may be contributing to the successful expansion in these areas.

Evidence for compensatory mutations in fluoroquinolone resistant strains of either *Shigella* species or *E. coli* remains limited[48], suggesting that the fitness costs associated with harbouring resistance to these drugs will be negligible. This result supports evidence in previous studies have suggesting that particular quinolone-resistance conferring SNPs, notably in *gyrA*, will not be associated with any loss of fitness in some organisms[49].

Unfortunately, a limitation in this study was the lack of phenotypic data on resistance for a range of antibiotics and, though *in silico* testing allowed for classification of susceptible/resistant to fluoroquinolones, full phenotypic testing would extend the scope of the results in this chapter. In spite of these challenges, the analysis undertaken here identified some clear differences between antibiotic resistance profiles in Vietnamese *Shigella* populations that may be contributing to the replacement of *S. flexneri* with *S. sonnei* in this region.

References

1. Blair, J. M. A., Webber, M. A., Baylay, A. J., Ogbolu, D. O. & Piddock, L. J. V. Molecular mechanisms of antibiotic resistance. *Nat. Rev. Microbiol.* **13,** 42–51 (2015).

2. Courvalin, P. Why is antibiotic resistance a deadly emerging disease? *Clin. Microbiol. Infect.* **22,** 1–3 (2016).

3. Laxminarayan, R. *et al.* Antibiotic resistance-the need for global solutions. *Lancet Infect. Dis.* **13,** 1057–1098 (2013).

4. Andersson, D. I. & Hughes, D. Antibiotic resistance and its cost: is it possible to reverse resistance? *Nat. Rev. Microbiol.* **8,** 260–71 (2010).

5. Maisnier-Patin, S. & Andersson, D. I. Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. *Res. Microbiol.* **155,** 360–9 (2004).

6. Maisnier-Patin, S., Berg, O. G., Liljas, L. & Andersson, D. I. Compensatory adaptation to the deleterious effect of antibiotic resistance in Salmonella typhimurium. *Mol. Microbiol.* **46,** 355–66 (2002).

7. Holt, K. E. *et al.* Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44,** 1056–9 (2012).

8. Shiferaw, B. *et al.* Antimicrobial susceptibility patterns of shigella isolates in foodborne diseases active surveillance network (foodnet) sites, 2000-2010. *Clin. Infect. Dis.* **54,** (2012).

9. Holt, K. E. *et al.* Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17522–7 (2013).

10. Konradt, C. *et al.* The Shigella flexneri type three secretion system effector IpgD inhibits T cell migration by manipulating host phosphoinositide metabolism. *Cell Host Microbe* **9,** 263–272 (2011).

11. Connor, T. R. *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri. *Elife* **4,** 1–16 (2015).

12. Ventola, C. L. The antibiotic resistance crisis: part 1: causes and threats. *P T A peer-reviewed J. Formul. Manag.* **40,** 277–83 (2015).

13. Comas, I. *et al.* Whole-genome sequencing of rifampicin-resistant Mycobacterium tuberculosis strains identifies compensatory mutations in RNA polymerase genes. *Nat. Genet.* **44,** 106–10 (2012).

14. DiNardo, S., Voelkel, K. a, Sternglanz, R., Reynolds, a E. & Wright, a. Escherichia coli DNA topoisomerase I mutants have compensatory mutations at or near DNA gyrase genes. *Cold Spring Harb. Symp. Quant. Biol.* **47 Pt 2,** 779–84 (1983).

15. Usongo, V., Tanguay, C., Nolent, F., Bessong, J. E. & Drolet, M. Interplay between type 1A topoisomerases and gyrase in chromosome segregation in Escherichia coli. *J. Bacteriol.* **195,** 1758–68 (2013).

16. Casali, N. *et al.* Microevolution of extensively drug-resistant tuberculosis in Russia. *Genome Res.* **22,** 735–45 (2012).

17. Szamecz, B. *et al.* The Genomic Landscape of Compensatory Evolution. *PLoS Biol.*

12, (2014).

18. Kryazhimskiy, S. & Plotkin, J. B. The population genetics of dN/dS. *PLoS Genet.* **4,** (2008).

19. Farhat, M. R. *et al.* Genomic analysis identifies targets of convergent positive selection in drug-resistant Mycobacterium tuberculosis. *Nat. Genet.* **45,** 1183–9 (2013).

20. Cui, Y. *et al.* Historical variations in mutation rate in an epidemic pathogen, Yersinia pestis. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 577–82 (2013).

21. Wellcome Trust Sanger Institute. No Title. (2016). at <http://www.sanger.ac.uk/resources/downloads/bacteria/shigella.html>

22. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30,** 2114–2120 (2014).

23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754–60 (2009).

24. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

25. Stamatakis, A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).

26. Paradis, E., Claude, J. & Strimmer, K. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics* **20,** 289–290 (2004).

27. Schliep, K. P. phangorn: Phylogenetic analysis in R. *Bioinformatics* **27,** 592–593 (2011).

28. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33,** 284–288 (2005).

29. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *bioRxiv* 118000 (2017). doi:10.1101/118000

30. Jombart, T. Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics* **24,** 1403–1405 (2008).

31. Marcusson, L. L., Frimodt-Møller, N. & Hughes, D. Interplay in the selection of fluoroquinolone resistance and bacterial fitness. *PLoS Pathog.* **5,** e1000541 (2009).

32. Plamann, M. D. & Stauffer, G. V. Regulation of the Escherichia coli glyA gene by the metR gene product and homocysteine. *J. Bacteriol.* **171,** 4958–62 (1989).

33. Thompson, C. N. *et al.* Clinical implications of reduced susceptibility to fluoroquinolones in paediatric Shigella sonnei and Shigella flexneri infections. *J. Antimicrob. Chemother.* **71,** 807–815 (2016).

34. Chatterjee, P. K. & Sternberg, N. L. A general genetic approach in Escherichia coli for determining the mechanism(s) of action of tumoricidal agents: application to DMP 840, a tumoricidal agent. *Proc Natl Acad Sci U S A* **92,** 8950–8954 (1995).

35. Duval, V. & Lister, I. M. MarA, SoxS and Rob of Escherichia coli – Global regulators of multidrug resistance, virulence and stress response. *Int. J. Biotechnol. wellness Ind.*

2, 101–124 (2013).

36. Liu, B. T. *et al.* Detection of mutations in the gyrA and parC genes in Escherichia coli isolates carrying plasmid-mediated quinolone resistance genes from diseased food-producing animals. *J. Med. Microbiol.* **61,** 1591–1599 (2012).

37. Lázár, V. *et al.* Genome-wide analysis captures the determinants of the antibiotic cross-resistance interaction network. *Nat. Commun.* **5,** 4352 (2014).

38. Minarini, L. A. R. & Darini, A. L. C. Mutations in the quinolone resistance-determining regions of gyrA and parC in enterobacteriaceae isolates from Brazil. *Brazilian J. Microbiol.* **43,** 1309–1314 (2012).

39. Sulavik, M. C., Gambino, L. F. & Miller, P. F. The MarR repressor of the multiple antibiotic resistance (mar) operon in Escherichia coli: prototypic member of a family of bacterial regulatory proteins involved in sensing phenolic compounds. *Mol. Med.* **1,** 436–46 (1995).

40. Grove, A. MarR family transcription factors. *Curr. Biol.* **23,** R142–R143 (2013).

41. Kotloff, K. L. *et al.* Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull. World Health Organ.* **77,** 651–66 (1999).

42. Mansilla, M. C. & Mendoza, D. De. The Bacillus subtilis cysP gene encodes a novel sulphate permease related to the inorganic phosphate transporter (Pit) family. 815–821 (2016).

43. Caldwell, A. M. & Smith, R. L. Membrane Topology of the ZntB Efflux System of Salmonella enterica Serovar Typhimurium. **185,** 374–376 (2003).

44. Divya, M. P., Mathew, P. D., Jyothi, R., Bai, R. & Thomas, S. Mutations in gyrA & parC genes of Shigella flexneri 2a determining the fluoroquinolone resistance. *Indian J. Med. Res.* **141,** 836–838 (2015).

45. Bagel, S., Hüllen, V., Wiedemann, B. & Heisig, P. Impact of gyrA and parC mutations on quinolone resistance, doubling time, and supercoiling degree of Escherichia coli. *Antimicrob. Agents Chemother.* **43,** 868–875 (1999).

46. Stupina, V. a & Wang, J. C. Viability of Escherichia coli topA mutants lacking DNA topoisomerase I. *J. Biol. Chem.* **280,** 355–60 (2005).

47. Baker, S. *et al.* Fitness benefits in fluoroquinolone-resistant Salmonella Typhi in the absence of antimicrobial pressure. *Elife* **2,** e01229 (2013).

48. Kishii, R. & Takei, M. Relationship between the expression of ompF and quinolone resistance in Escherichia coli. *J. Infect. Chemother.* **15,** 361–366 (2009).

49. Rozen, D. E., McGee, L., Levin, B. R. & Klugman, K. P. Fitness costs of fluoroquinolone resistance in Streptococcus pneumoniae. *Antimicrob. Agents Chemother.* **51,** 412–416 (2007).

Chapter 5

Environmental persistence and niche adaptation in Vietnamese *Shigella sonnei* and *Shigella flexneri* – insights from the pan genome

5.1 Abstract

Advances in next-generation sequencing (NGS) have driven a range of studies to characterise the key genomic features and epidemiology of a variety of important human and wildlife diseases. The spread of an infectious pathogen can be driven by its relative fitness, and this is often influenced by its ability to adapt to selective pressures, such as exposure to antimicrobials and changes to the host environment. Bacteria, notably, can acquire genes from other organisms or the environment through horizontal gene transfer and natural transformation, allowing for fast adaptation to challenges and rapid fixation of beneficial genes. Such selective pressures can contribute to variation in gene content between lineages, with genes overrepresented in certain lineages constituting likely candidates for recent or ongoing adaptation.

In this study, I have used *de novo* assembly to construct the pan genome of *Shigella sonnei* and *Shigella flexneri* in Vietnam, the two most common species causing shigellosis, or bacillary dysentery. Though *S. flexneri* has historically been the most prevalent species in developing countries, including Vietnam, there has been an observed shift towards infections caused by *S. sonnei*.

Our results show variation in the gene content of the average pan genomes of these *Shigella* species, with a number of genes identified in *S. sonnei* that can be associated with metabolism and host adaptation. These key phenotypic differences between *Shigella* species may influence survival and niche adaptation in a changing environment, and could be one of the factors driving the spread of *S. sonnei* in many developing countries.

5.2 Introduction

Whole genome sequencing (WGS) of common bacterial species has now become relatively inexpensive, with an abundance of new research aiming to improve technologies that increase the efficiency and accuracy of genome assemblies. As such, our knowledge of the genetic variation and evolution of major disease causing bacteria, such as *Escherichia coli*[1–4] and *Staphylococcus aureus*[5] is very well understood, with numerous studies aiming to characterise the genes that are present in these species, as well as their functions. Applying this knowledge to the epidemiology and transmission of emerging or re-emerging pathogens can help to identify genetic drivers that may be responsible for influencing the spread of a disease in a particular environment[6–8].

Short read sequencing is the most commonly utilised technology for WGS due to its relatively low cost; piecing together these short, fragmented stretches of DNA to reconstruct the 'true' complete genome of a species does, though, present some challenges[9]. The standard method for building genomes and detecting variation among samples of previously characterised species is to map the short reads of sequenced DNA against a previously assembled, closely related 'reference' strain[10]. This will allow comparison of the similarity of shared genes between the reference and sampled strains by identifying single nucleotide polymorphisms (SNPs), single sites of variation, which can provide insights into their homology and evolution. This method is often favourable as it is generally fairly straightforward to assemble a large number of genomes with reasonable accuracy. The major shortfall though is that only genomic regions that are present in the reference strain will be mapped in the sampled strains, so the full genetic variation between samples will not be captured[9]. Additionally, regions missing from any one sample will be ignored when conducting phylogenetic reconstructions from alignments and so information on genomic regions present in the reference strains may also be lost if these are not shared by all strains. This is somewhat problematic in bacterial taxa that can undergo high rates of horizontal gene transfer (HGT) or in cases where a closely related reference genome is not available[11,12].

In contrast, the *de novo* assembly of genomes involves stitching together the fragmented short reads of DNA without the use of a reference strain, instead relying on complex computations and probabilistic methods to construct continuous stretches of DNA, or contigs[9]. This process commonly uses De Bruijn graphs to construct a putative consensus sequence through overlapping reads by a particular length, the *k* parameter or hash length, of the matching nucleotides[13]. These uninterrupted 'contigs' can contain multiple genes including those that may not have been present in a reference strain. Contigs can also then

be joined and extended by 'scaffolding' with reads mapped to a reference to capture any genetic information which may be present in the reference strain but were missed through inaccuracies in the *de novo* assembly. The contigs can be annotated to determine protein coding regions which can then compared to the sequences of previously described genes in other species or strains through tools such as BLAST[14].

Due to the plasticity of the bacterial genome through processes such as HGT and conjugation, it is thought that the 'true' complete genome of a species can never be fully described[15]. However, reconstructing genomes with *de novo* assemblies can result in a greater approximation of content of both the 'core genome' - genes that are shared by all individuals – and 'accessory genome' - genes found at different frequencies (<99%) – of a population. The core and accessory genomes can be described together as the 'pan-genome' of a particular species. Describing the differences in pan genomes either between or within species allows us to use the presence and absence of genes as the unit of comparison, as well as the variation within core genes from standard reference-based SNP analysis[12,15–19]. This is particularly effective for discovering new genes and functions or identifying recombination and gene loss events within in a population.

Shigella is a genus of Gram-negative bacteria responsible for causing shigellosis, or bacillary dysentery, which can cause severe diarrhoea and even death in some cases, particularly in children under 5 years of age or people with a compromised immune response[20,21]. There are four serogroups that have formed the basis for the delineation into species: *S. sonnei, S. flexneri, S. boydii* and *S. dysenteriae,* with all but *S. sonnei* comprising multiple serotypes based on the somatic O-antigen, and the serocomplex is closely related to *Escherichia coli*[22,23]. The most commonly found species are *S. sonnei* and *S. flexneri* with the latter historically associated with infections isolated from developing countries[24–27]. These species also show different epidemiological patterns, with *S. flexneri* typically causing a more severe disease and being responsible for more fatalities[28].

Recently, patterns have begun to emerge of a species shift in many developing countries, with shigellosis infections caused by *S. sonnei* becoming more prevalent to the degree that it is now replacing *S. flexneri* as the primary agent of bacillary dysentery in many of these regions[28–30]. The emergence of *S. sonnei* in these regions has been investigated using phylogeographic tools [31,32] and different studies have identified differences in transmission[33,34], epidemiological[35–37] and antimicrobial resistance[26,30,38] features between the two species. Yet, the reasons for the species replacement remain unclear as there is some uncertainty over how these factors may have contributed to the observed spread of *S. sonnei.* Additionally, there is evidence that there may be an environmental component in the

transmission of *Shigella spp.,* with organisms able to occupy environmental reservoirs such as on food[37,39,40], fomites[41] and in waterways[24,29,42], though again their impact on the transmission and potential differences between *Shigella* species is not yet fully understood.

Direct and indirect associations with other organisms both in the environment and within a host may also impact on the relative shifts in distribution of the two species. *S. sonnei* has been shown to be protected in the environment through phagocytosis by the common amoeba *Acanthamoeba castellanii* which can shield the bacteria from otherwise inhabitable conditions[43,44]. There is also evidence of secondary host immunity to *S. sonnei* acquired through exposure to *Plesiomonas shigelloides*, a bacterial species commonly occupying poor quality water systems, which shares the same major lipopolysaccharide (LPS) O-side chain surface antigen, the first target of an immune response[22,45].

Both *S. flexneri* and *S. sonnei* contain a major plasmid, pINV, which encodes many essential and beneficial phenotypic traits, including virulence and antibiotic resistance[46]. This plasmid is less likely to be maintained in environmental temperatures in *S. sonnei* though, which will lower the ability of the species to cause disease outside of the host[47]. This may suggest key environmental differences between *Shigella* species, with *S. sonnei* better adapted to life within a host with transmission as person-to-person.

Despite the previous efforts to identify the causes behind the replacement in the two Shigella species, the possible role of differences in genome content between the two species has been ignored to date. This prompted a study aiming to reconstruct the complete pan genome of clinical strains of *S. sonnei* and *S. flexneri* isolated from Vietnam, a region that has recently undergone this pattern of species replacement. I expect that by using *de novo* assembly rather than a reference-based approach I will be able to better capture the full genomic content of both *Shigella* species in this population and identify genes that may contribute to phenotypic and ecological differences. In addition, a set of globally distributed strains were analysed to characterize broad species-level variation in gene content. Additionally, understanding the genetics underlying differences in *Shigella* spp. Life histories and epidemiology may help us to improve the management of shigellosis infections, and more generally contribute to our knowledge of infectious disease emergence and transmission[27].

5.3 Methodology

Shigella isolates

135 *Shigella flexneri* and 146 *Shigella sonnei* clinical isolates from three regions in Vietnam, Ho Chi Minh, Hue and Khanh Hoa, were selected from data sequenced in Holt *et al*., 2013[32], with raw paired-read FASTQ files obtained from the Wellcome Trust Sanger Institute online database[48]. In addition, 60 isolates of both species were chosen from a global dataset comprising *S. sonnei* and *S. flexneri* samples from 25 and 34 countries respectively, with all subsequent analysis carried out separately to the Vietnamese samples for later pan genome comparison. Sequence data was also taken from the Wellcome Trust Sanger Institute online database[48], details of the selected samples including accession number and location are available in the **supplementary materials.**

De Novo assembly

Paired read FASTQ files were quality checked with adapter removal and read trimming was carried out where required using the methods detailed in Chapter 4. High quality forward and reverse reads were assembled for each sample strain using VelvetOptimiser[49], a Perl wrapper tool for optimising Velvet assemblies. Velvet attempts to build contigs, continuous lengths of uninterrupted nucleotides, by constructing a De Bruijn graph that works to link lengths of matching of nucleotide bases. These lengths of nucleotides are substrings of the reads called *k*-mers with a length determined by the hash/word length or *k* parameter, and these will be linked by *k*-1 bases so that a contig is extended by one base with each added *k*-mer. VelvetOptimister optimizes the hash/word length and expected coverage parameters in the assembly by searching through user defined minimum and maximum bounds. Assembled contigs were then joined by scaffolding with raw reads mapped to complete genomes of a reference strain (*S. sonnei* strain 046 NC_007384 or *S. flexneri* strain 2a 301 NC_004337.2) using AlignGraph[50]. Both extended and remaining contigs were then combined into one FASTA file for annotation and building the pan genome.

Annotation and pan genome construction

Prokka[51] was used to annotate assembled and joined contigs (>100bp) with genomic features such as genes and RNAs as well as hypothetical coding regions which do not match any previously annotated features found in the searched databases. The resulting

annotated contig files were then analysed by Roary[52] to build the pan genome and resulting presence/absence tables of homologous gene clusters (HGCs) through clustering of transcribed protein sequences by similarity (98% CD-HIT, 95% BLASTP sequence identity). Roary assigns a gene name and genomic function annotation to each HGC where possible through searches with BLASTP, with paralogous genes (gene duplication events) identified and retained in separate clusters. Separate HGCs may also be assigned the same gene name and function in cases where the BLASTP result is the same gene by descent from a different species of origin (orthology). The results were separated into isolates belonging to either *S. flexneri* or *S. sonnei* and the presence/absence of HGCs in these samples recalculated in *R*[53]. When analysing homologous groups, hypothetical proteins and duplicated genes that were markedly shorter than the annotated gene from the origin species were removed, as these are often indicative of non-functional pseudogenes. The complete workflow is described in **Figure 1**.

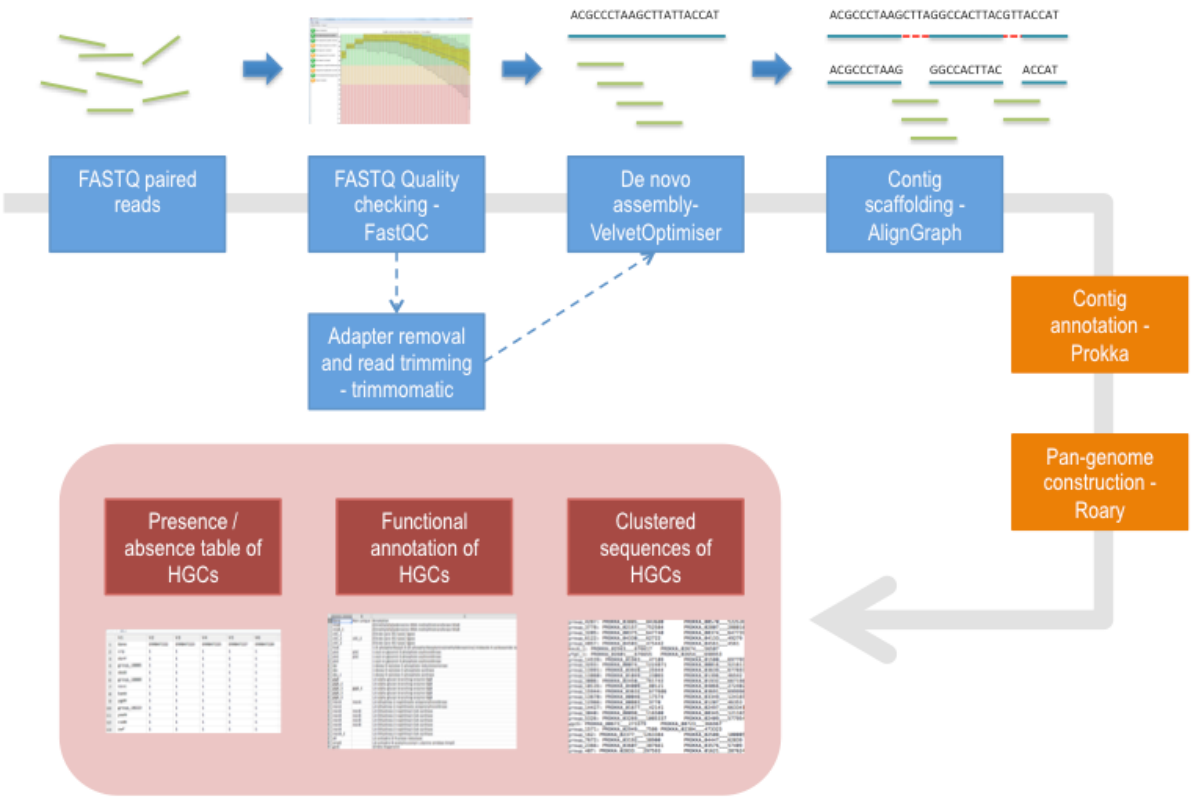


Figure 1: A methodological workflow of *de novo* assembly and pan genome construction as used in this study.

Analysis of homologous groups

Pairwise all-against-all genetic distance of the consensus contig sequences of all isolates was carried out using MASH[54] which calculates simple, rapid distance estimation between sequences using a MinHash algorithm. The resulting distance matrix was compared to a matrix of the pairwise shared number of HGCs between isolates by calculating the Pearson product-moment correlation coefficient to determine whether there is a significant correlation between gene presence/absence and genome wide diversity (including within gene variation and non-coding regions).

Gene ontology (GO) terms were found by clustering annotated gene names into categories of molecular function with the PANTHER Classification System from the Gene Ontology Consortium[55,56] to illustrate broad, genome-wide patterns of gene function predicted for each species. Specific HGCs assigned a gene name that were present in significantly different frequencies in each species were manually reviewed for previous characterization of the function and resulting phenotype in *Shigella* or related species.

	Shigella flexneri		***Shigella sonnei***	
	HGCs	**HGCs w/ assigned gene**	**HGCs**	**HGCs w/ assigned gene**
Core (99% <= isolates <= 100%)	1928	1902	2951	2722
Softcore (95% <= isolates < 99%)	445	560	475	462
Shell (15% <= isolates < 95%)	2994	1432	1495	559
Cloud (0% <= isolates < 15%)	7683	2157	4988	1280
Total	13050	6051	9909	5023

Table 1: The number of Homologous Gene Clusters (HGCs) in *S. sonnei* and *S. flexneri* both before and after the removal of potential pseudogenes, hypothetical proteins and non-annotated groups.

5.4 Results

Pan genome estimation

Overall, 12,708 homologous gene cluster (HGCs) were identified through the construction of the *de novo* pan genome of all isolates belonging to both *S. sonnei* and *S. flexneri*. The removal of groups missing a gene name and putative functional annotation (hypothetical proteins) decreased the number to 6,854 clusters and of the annotated HGCs there were 3,003 unique previously described gene names. The average number of annotated HGCs identified for a single sample is 3,548 (σ=61.66) in *S. sonnei* and 3,417 (S σ=78.9) in *S. flexneri*. This is in line with previous estimates for the number of CDS regions in the reference genomes for each species, with 3,569 genes in *S. sonnei* 046 and 3,565 in *S. flexneri* 301[21].

The construction of the pan genome from *de novo* assemblies identified 1,104 genes found in *S. sonnei* (916 in >=95% of isolates) and 1,237 in *S. flexneri* (718 in >=95% of isolates) that were not found in the respective *S. sonnei* 046 and *S. flexneri* 301 reference genomes. This reflects the variability in gene content that can be present between tested strains and widely used reference genomes thus underlining the advantage of using *de novo* assemblies when searching for novel genes or functions.

The pan genome of *S. sonnei* (with hypothetical proteins and potential pseudogenes removed) comprises 5,023 HGCs, with 2,722 HGCs classed as 'core', found in >=99% isolates, and 462 HGCs 'softcore', >=95% isolates (**Table 1**). The pan genome of *S. flexneri* differs in its relative proportion of the 6,051 total HGCs classified as core and softcore, with 1,902 and 560 HGCs in each category respectively. This number is again analogous to previous estimates of the core genome of *E. coli* and *Shigella* spp.[2], although the number of core genes in *S. sonnei* is marginally higher, likely owing to the strongly clonal evolution of the tested isolates[32]. The number of rare HGCs ('cloud' genes, shared by <15% isolates) was significantly higher in the *S. flexneri* pan genome with these clusters accounting for around 35.65% of the total HGCs identified in *S. flexneri* samples compared to 25.58% in *S. sonnei*.

The difference in the number of shared HGCs found in at least 95% of strains between *S. sonnei* and *S. flexneri* may reflect the true difference in softcore genome size or it may be biased through variation in intra-species genetic diversity. It has been reported that Increasing the number of samples and adding more diverse strains when estimating bacterial pan genomes will decrease the number of genes shared between isolates and reduce the size of the core genome[19,57].

Rarefaction curves were produced for both the shared and total number of HGCs for each species by randomly sampling an increasing number of isolates, (up to the total number for each species) over 1,000 iterations, and estimating the number of total and shared genes captured at these sample numbers (**Figure 2**). The total HGC accumulation curves (**Figure 2A**) for both *S. sonnei* and *S. flexneri* begin to plateau at around 35 samples, with 95% of the total HGCs captured, suggesting that adding further samples will only negligibly increase the number of novel HGCs identified and thus the number of isolates used in this study accurately capture the majority of the total genes found in the population. Consistent with the patterns shown in previous studies, the rarefaction curves for the number of shared HGCs decline as more samples are added (**Figure 2B**), indicating that the core genome of a species will also decrease. The number of shared genes in *S. sonnei,* although not reaching a clear plateau, appears to decline less rapidly than in *S. flexneri* indicating fewer shared genes between any two *S. flexneri* samples on average, and suggesting higher diversity within the *S. flexneri* isolates included in this study.

To determine whether genetic diversity was linked to the number of shared HGCs, and thus the size of the core genome, I computed the pairwise genetic distance between samples and compared these estimates to the number of shared genes between each pair using the Pearson product-moment correlation coefficient (**Figure 3**). The average intra specific genetic distance between pairs was 1.175×10^{-3} ($\sigma = 5.32 \times 10^{-4}$) for *S. sonnei* samples and 6.62×10^{-3} ($\sigma = 6.38 \times 10^{-3}$) for *S. flexneri*, with the average number of shared genes 6,562.77 (σ=88.879) and 6,193.78 (σ=347.67) respectively. Overall there was a significant negative correlation between genetic diversity and number of shared genes (Pearson's r = -0.9674). As the genetic diversity of *S. flexneri* was higher than *S. sonnei* this may explain the reason that the number of shared HGCs was lower in this species and the greater number of novel and low frequency HGCs identified.

Genome wide functional variation

Unique gene names assigned to HCGs identified in each species, both overall and in at least the softcore genome, were analysed with the PANTHER Classification System to infer potential functional differences between *S. sonnei* and *S. flexneri* (**Table 2)**. The categories used for classification were based on GO-slim terms for molecular function to capture a general representation of the broad function of the genes in each group. Of the 2,839 unique gene names identified in the total pan genomes of *S. sonnei* and 2,921 in *S. flexneri*, 1,720 and 1,748 genes were assigned to a functional category. The most represented category was for both species was catalytic activity (GO:0003824), genes involved in catalysis of

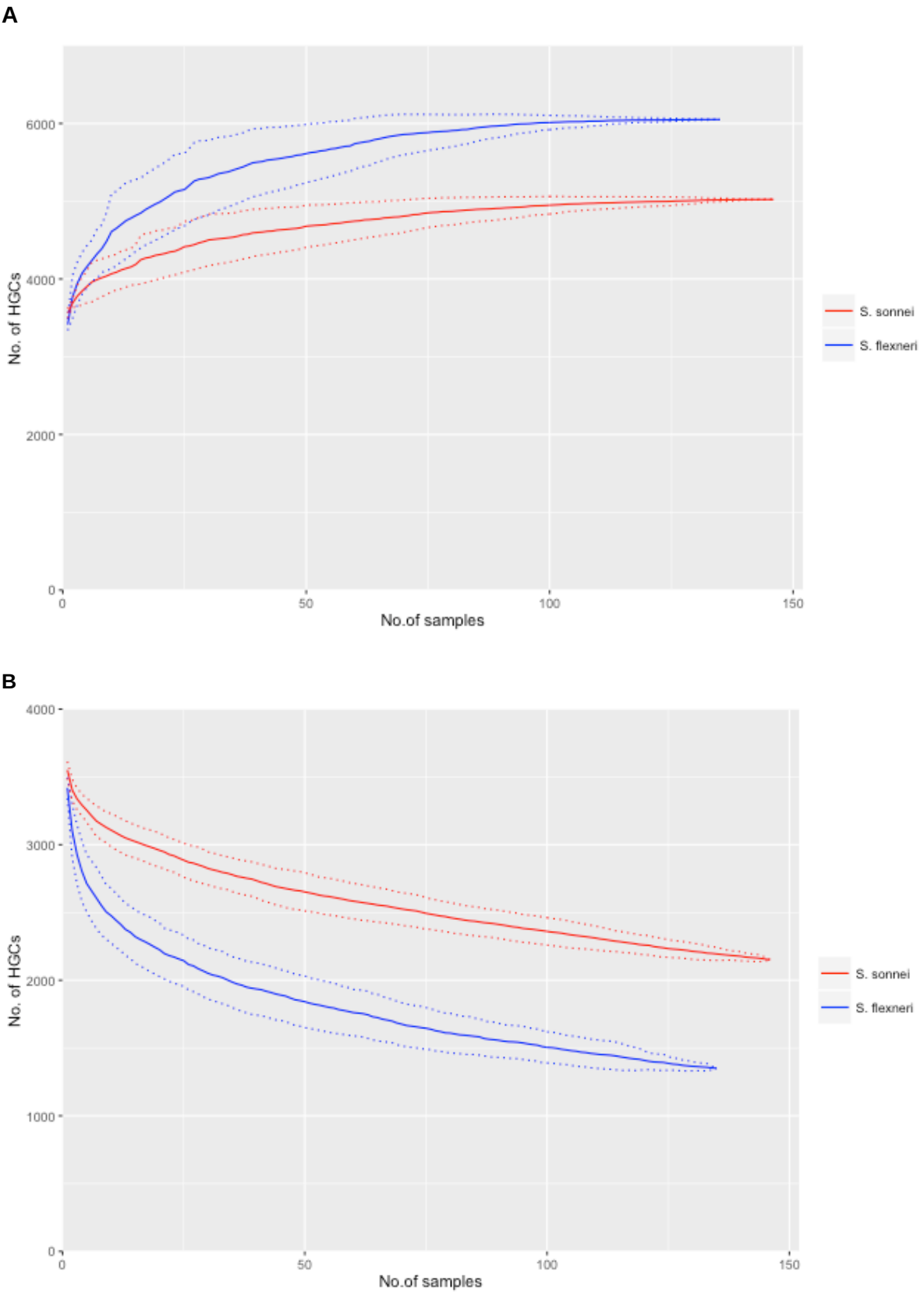


Figure 2: a) Accummulation and **b)** rarefaction curves illustrating the number of new and shared HGCs that were identified within each species when increasing the number of samples. Random sampling was carried out 1000 times at each sampling number to attain an average number of HGCs. Dotted lines represent the upper and lower bounds of HGCs present and solid line is the mean average.

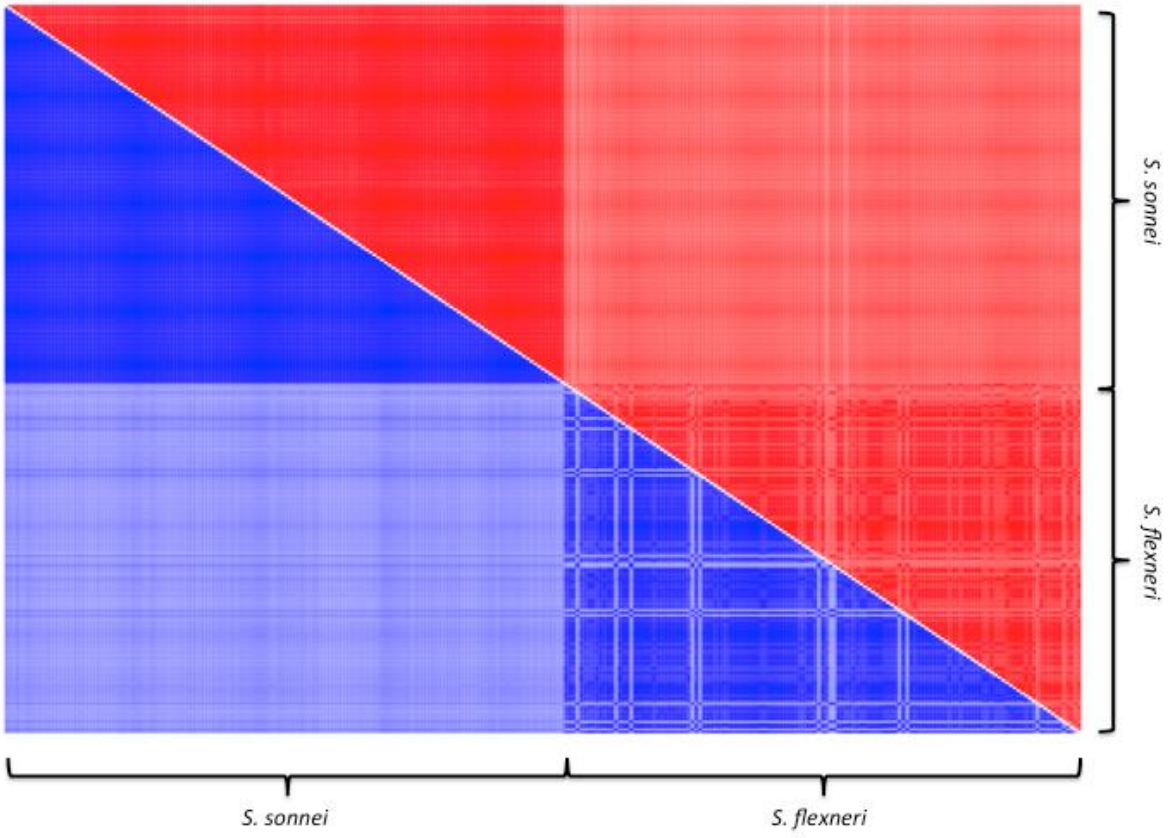


Figure 3: A heatmap illustrating the number of shared genes (red) and the genetic distance (blue) between all pairs of isolates used in this study. Genetic distance was measured by a MinHash algorithm and correlation between the number of shared genes and distance was computed using the Pearson product-moment correlation coefficient.

substrates in biochemical reactions, with the majority of other genes involved in binding (GO:0005488) and transporter activity (GO:0005215).

Overall, there was no significant difference between numbers of genes represented by each molecular function category in each species ($X^2 = 2.080$, df = 11, p-value = 0.9982); this was also true for genes found in at least 95% of isolates, the softcore genome, ($X^2 = 3.734$, df = 11, p-value = 0.9771). The *bfr* gene found to be present in all sampled *S. flexneri* genomes and allocated the molecular function of nutrient reservoir activity (GO:0045735), was the only instance where a functional category was present in only one of the species. This gene confers bacterioferritin, a haemoprotein involved in iron storage, which has been linked to oxidative stress response in *Bacteroides fragilis,* an anaerobic bacterial pathogen, with mutants lacking this gene exhibiting a lower oxygen tolerance[58,59].

	Number of genes		% of total genes	
	S. sonnei	***S. flexneri***	***S. sonnei***	***S. flexneri***
Catalytic activity (GO:0003824)	1015 (947)	1027 (835)	41.1% (40.7%)	40.8% (42.1%)
Binding (GO:0005488)	279 (258)	282 (238)	11.3% (11.1%)	11.2% (12.0%)
Transporter activity (GO:0005215)	224 (211)	231 (175)	9.1% (9.1%)	9.2% (8.8%)
Nucleic acid binding transcription factor activity (GO:0001071)	70 (66)	70 (56)	2.8% (2.8%)	2.8% (2.8%)
Structural molecule activity (GO:0005198)	58 (56)	58 (54)	2.3% (2.4%)	2.3% (2.7%)
Translation regulator activity (GO:0045182)	37 (31)	40 (31)	1.5% (1.5%)	1.6% (1.6%)
Antioxidant activity (GO:0016209)	12 (12)	12 (10)	0.5% (0.5%)	0.5% (0.5%)
Enzyme regulator activity (GO:0030234)	11 (9)	14 (8)	0.4% (0.4%)	0.6% (0.4%)
Electron carrier activity (GO:0009055)	5 (4)	5 (4)	0.2% (0.2%)	0.2% (0.2%)
Protein binding transcription factor activity (GO:0000988)	5 (5)	4 (4)	0.2% (0.2%)	0.2% (0.2%)
Receptor activity (GO:0004872)	4 (4)	4 (3)	0.2% (0.2%)	0.2% (0.2%)
Nutrient reservoir activity (GO:0045735)	0 (0)	1 (1)	0.0% (0.0%)	0.0% (0.1%)

Table 2: Classification of HGCs by Molecular Function GO term by the PANTHER Functional Classification System.

Species-specific genome content

Comparing the presence and absence of specific genes in the pan genomes of *S. sonnei* and *S. flexneri* allows for the inference of potential differences in expressed phenotypes between the species. Notably, genes found in essentially every sample from one species (in the core or softcore genomes) yet at very low frequencies or absent in the other can point towards genes that characterise true functional differences between the species. In addition, due to the relatively high relatedness between *Shigella spp.,* a gene that is completely absent in one species yet in the core or softcore of another suggests either a gain of a gene function and subsequently sweep to near or complete fixation, or can be evidence of gene loss through genetic drift or species specific selection differences. Either mechanism will

shape the pan genome content and potential functional variation that may be important in informing the ecology or lifestyle of an organism.

There were 117 HGCs present in at least the core genome of *S. sonnei* (>99% of isolates) that were not found in any of the pan genomes of *S. flexneri* isolates and this number rises to 131 when including those in the softcore genome of *S. sonnei* (>=95% of isolates). There were fewer HGCs present only in the *S. flexneri* pan genome, with 34 unique HGCs in the core and 53 in the softcore genome likely due to the smaller core and softcore genome compared to *S. sonnei*. These numbers include separate HGCs assigned the same gene name and those that were not assigned a gene name but were given a putative function, typically integrated phage proteins. These genomic features still may confer some phenotypic variation in the species that could influence the ecology.

Expanding the analysis to include HGCs that were present in one species at a high frequency (softcore genome, >=95% isolates) but were also present at low frequencies in the other species (<15% isolates) further identifies genes that may be likely candidates for broad functional differences between the species. An additional 185 HGCs were found in at least the softcore genome of *S. sonnei* that were found in less than 15% of *S. flexneri* isolates, and 12 HGCs in the softcore of *S. flexneri* that were found at a low frequency in *S. sonnei*. Again, the difference in the number of HGCs found in these proportions between the species is likely a result of the larger core genome of *S. sonnei* as well as the relatively smaller number of cloud, or low frequency, HGCs.

A full list of HGCs that were present, either solely or significantly more frequently, in each species is available in **Supplementary materials**.

Functional characterization of species-specific genes

Further analysis of the functional annotations assigned to HGCs that were present in significantly different frequencies in *S. sonnei* and *S. flexneri* revealed some putative candidate genes that may be linked to ecological and phenotypic differences between species (**Table 3)**. HGCs found in significantly higher numbers in only one species but with another HGC with the same assigned gene name and functional annotation present in the other species in more than 15% of isolates was excluded from this analysis, along with groups not assigned a gene name. Although there may be functional differences conferred by these features, this is beyond the scope of this study and it would require experimental validation to test for any variation in the expressed phenotype.

Of the genes identified that segregated either completely or in significantly higher numbers in *S. sonnei*, a large number have been associated with ecological niche adaptation, particularly through utilization of a variety of sources of catabolic substrates and stress response mechanisms. These associations include adaptations to the host organism, such as the *hcpA* gene, linked to intestinal adherence, and the platelet binding gene *gspB* previously isolated in *Streptococcus gordonii*, as well as to the external environment, including involvement in the nitroduction cycle (*nar* genes) and degradation of the herbicide atrazine in *Pseudomonas* with the gene *atzC*. The *S. flexneri* genome notably contained no genes that could be linked directly to environmental survival that were not also found in significant numbers of *S. sonnei.*

Genes involved in the catabolism of specific carbohydrates and other metabolic substrates were identified in markedly different proportions in *S. sonnei* and *S. flexneri*. Metabolic genes have been previously shown to distinguish between *E. coli* and *Shigella* genomes, with a number of genes found in *E. coli* absent in *Shigella spp*., proposed to be due to narrower niche specialization in *Shigella*[60]. The variation in metabolic mechanisms in *S. sonnei* and *S. flexneri* may demonstrate that these species are optimally adapted to employ different niches, with the larger catabolism associated genome content of *S. sonnei* appearing to enable the organism to utilize a wider breadth of energy sources.

Six genes (the *cas* operon and *ygb* gene family) associated with the CRISPR/Cas system of protection from phages and other mobile genetic elements were discovered in more than 95% of *S. sonnei* isolates, highlighting the presence of an acquired immunity mechanism in *S. sonnei* that appears not to be present in all the *S. flexneri* isolates sampled. This system may assist in the persistence and survival of this species through protection against bacteriophages, both in the host organism and the external environment[61].

The *S. sonnei* genome also contains a variety of genes linked to toxin/antitoxin systems that are not found in *S. flexneri*. These systems are common in prokaryotes and when found on plasmids are predominately linked to ensuring these plasmids are inherited by daughter cells after segregation[62]. Chromosomal toxin/antitoxin systems are less well understood and have been hypothesized to be associated with mechanisms for biofilm formation[63], antibiotic resistance[64] and gene regulation[65]. Interestingly, it has been theorized that these systems can be a response to nutrient stress or starvation, with the *E. coli MazEF* and *RelE2* toxin/antitoxin systems, identified here in *S. sonnei*, either inducing death in an individual to altruistically lower the nutrient uptake of the population, or by inhibiting translation to decrease individual nutrient demand[66,67]. These response mechanisms to nutrient depletion

could play an important role in ensuring the survival of *S. sonnei* in nutrient poor environments.

Antibiotic resistance associated genes were also identified in *S. sonnei*, including the *sbmC* gene that has been shown to be up-regulated in *Salmonella enterica var. Typhimurium* in the presence of sub-minimum inhibitory concentrations of fluroquinolones. The use of these antimicrobials is known to be widespread in Vietnam where bacterial resistance common[29].

Geographic gene differences

There were no HGCs that segregated completely by geographic region when considering both *S. sonnei* and *S. flexneri* samples together that were isolated from three provinces in Vietnam; Ho Chi Minh City, Hue and Khanh Hoa. In addition, no HGCs were significantly more likely to be found in both species isolated from one region (present in more than 95% of isolates compared to less than 15% of isolates from another province). This suggests that there is limited evidence of a shared gene pool and recombination of newly acquired genes between *S. sonnei* and *S. flexneri* in these local populations, and that the species-level pan genome has been formed predominately by species-specific gene acquisition and loss.

Comparing local populations of each *S. sonnei* and *S. flexneri* separately, there were also no HGCs that segregated by geography within each species, either completely or in markedly higher proportions. The majority of HGCs that were found only in one region were identified in single individuals, suggesting either that the majority of recombination is individual, or that it is transient within the population and the shared genome of each species will have been formed prior to the spread of the pathogen across Vietnam. Alternatively, it suggests that strongly beneficial genes will sweep rapidly to fixation or be acquired at multiple, convergent times within a species, such as has been shown previously with antimicrobial resistance genes in *S. sonnei* in this region[32]

HGC gene name	Function	Shigella sonnei presence (n=146)	Shigella *flexneri* presence (n=135)	Additional notes
aarA	Rhomboid protease AarA	146	13	Involved in aminoglycoside resistance in *Providencia stuartii*.[68]
abgA	p-aminobenzoyl-glutamate hydrolase subunit A	144	10	Folate uptake - increase resistance to sulfonamide antibiotics.[69]
abgB	p-aminobenzoyl-glutamate hydrolase subunit B	146	10	Folate uptake - increase resistance to sulfonamide antibiotics.[69]
abgT	p-aminobenzoyl-glutamate transport protein	146	10	Folate uptake - increase resistance to sulfonamide antibiotics.[69]
acoR	Acetoin dehydrogenase operon transcriptional activator AcoR	0	129	Involved in carbon storage or avoiding acidification. Acetoin excreted when with glucose or other fermentable carbon sources.[70]
adhB	Alcohol dehydrogenase 2	0	134	Ethanol oxidation in aerobic respiration.[71]
adhT	Alcohol dehydrogenase	0	133	Ethanol oxidation in aerobic respiration.[71]
aldB	Aldehyde dehydrogenase B	0	133	Ethanol oxidation in aerobic respiration.[72]
allA	Ureidoglycolate lyase	0	135	-
allB	Allantoinase	146	14	-
allR	HTH-type transcriptional repressor AllR	0	134	-
aplIR	Type-2 restriction enzyme AplI	146	0	-
arcC2	Carbamate kinase 2	146	0	-
argF	Ornithine carbamoyltransferase	145	13	-
atoA	Acetate CoA-transferase subunit beta	145	0	Degradation of short chain fatty acids.[73]
atoD	Acetate CoA-transferase subunit alpha	145	0	Degradation of short chain fatty acids.[73]
atoE	Short-chain fatty acids transporter	146	0	Degradation of short chain fatty acids.[73]
atzC	N-isopropylammelide isopropyl amidohydrolase	146	0	Degradation of herbicide Atrazine.[74]
avtA	Valine--pyruvate aminotransferase	0	134	-
bfd	Bacterioferritin-associated ferredoxin	0	135	Oxidative stress.

bfr	Bacterioferritin	0	135	-
cadA	Lysine decarboxylase, inducible	146	0	Bacterial acid stress.[75]
cadC	Transcriptional activator CadC	146	0	Bacterial acid stress.[75]
casA	CRISPR system Cascade subunit CasA	146	13	Protection against viruses and other mobile genetic elements.[76]
casC	CRISPR system Cascade subunit CasC	146	13	Protection against viruses and other mobile genetic elements.[76]
casD	CRISPR system Cascade subunit CasD	138	11	Protection against viruses and other mobile genetic elements.[76]
casE	CRISPR associated protein	144	13	Protection against viruses and other mobile genetic elements.[76]
ccmL	Carbon dioxide concentrating mechanism protein CcmL	146	0	Persistence in aquatic environments.[77]
cdhB	Caffeine dehydrogenase subunit beta	143	0	Degradation of caffeine in *Pseudomonas*.[78]
clpP1	ATP-dependent Clp protease proteolytic subunit 1	145	0	Acclimation to UV-B and low temperature in cyanobacteria.[79]
dacD	D-alanyl-D-alanine carboxypeptidase DacD precursor	146	0	Cell wall biosynthesis - penicillin binding protein (PBP).[80]
ddpA	putative D,D-dipeptide-binding periplasmic protein DdpA precursor	143	13	-
ddpB	putative D,D-dipeptide transport system permease protein DdpB	146	13	-
ddrA	Diol dehydratase-reactivating factor alpha subunit	146	19	Preserves genome integrity.[81]
dhaK	PTS-dependent dihydroxyacetone kinase, dihydroxyacetone-binding subunit DhaK	0	131	PTS-dependent dihydroxyacetone kinase instead of ATP - involved in carbohydrate uptake and control of carbon metabolism in bacteria.[82]
dhaL	PTS-dependent dihydroxyacetone kinase, ADP-binding subunit DhaL	0	130	PTS-dependent dihydroxyacetone kinase instead of ATP - involved in carbohydrate uptake and control of carbon metabolism in bacteria.[82]
dinJ	Antitoxin DinJ	146	0	With yafQ forms dinJ-yafQ operon - toxin/antitoxin stress response.[83]
dlgD	2,3-diketo-L-gulonate reductase	0	133	-
dmlA	D-malate dehydrogenase [decarboxylating]	146	13	Growth on D-malate as sole carbon source and in butanoate metabolism, a by-product of anaerobic respiration and highly concentrated in the human gut.[84]
ecfT	Energy-coupling factor transporter transmembrane protein EcfT	0	134	Cellular transport.

envR	putative acrEF/envCD operon repressor	146	13	Drug efflux pump gene.[85]
flr	Flavoredoxin	145	0	Only characterised *in Desulfovibrio gigas.*
frmA	S-(hydroxymethyl)glutathione dehydrogenase	144	0	Detoxificaiton of formaldehyde.[86]
frmB	S-formylglutathione hydrolase FrmB	144	0	Detoxificaiton of formaldehyde.[86]
frmR	Transcriptional repressor FrmR	144	0	Detoxificaiton of formaldehyde.[86]
galF	UTP--glucose-1-phosphate uridylyltransferase	0	135	Increasing the intracellular concentration of UDP-glucose and enhancing the thermal stability of the UDP-glucose pyrophosphorylase.[87]
gbpR	HTH-type transcriptional regulator GbpR	146	10	Found in *Agrobacterium* - involved in sugar chemotaxis.[88]
gno	Gluconate 5-dehydrogenase	146	0	Oxidation of glucose.[89]
gspB	Putative general secretion pathway protein B	146	0	Platelet binding in *Streptococcus gordonii*.[90]
guaD	Guanine deaminase	145	19	-
hcpA	Major exported protein	146	7	May be linked to intestinal adherence.[91]
hicB	Antitoxin HicB	146	8	Toxin/antoxin operon along with HicA - nutrient stress.[92]
hilA	Transcriptional regulator HilA	146	0	Invasion gene regulator in *Salmonella entericaserovar Typhimurium.*[93]
hipA	Serine/threonine-protein kinase HipA	146	0	Toxin/antitoxin system - mutations linked to multidrug resistant 'persister' cells and can influence biofilm formation.[94]
hipB	Antitoxin HipB	146	0	Toxin/antitoxin system - mutations linked to multidrug resistant 'persister' cells and can influence biofilm formation.[94]
hyuA	D-phenylhydantoinase	146	13	-
iadA	Isoaspartyl dipeptidase	0	132	-
idnD	L-idonate 5-dehydrogenase (NAD(P)( ))	146	1	Growth on L-idonate as sole carbon source.[95]
idnT	Gnt-II system L-idonate transporter	146	0	Growth on L-idonate as sole carbon source.[95]
iraD	Anti-adapter protein IraD	0	131	Stress response.[96]
lacY	Lactose permease	146	0	Lactose metabolism.[97]
leuE	Leucine efflux protein	146	13	-

lysU	Lysine--tRNA ligase, heat inducible	146	0	Heat shock/stress response.
malS	Alpha-amylase precursor	0	134	-
matA	HTH-type transcriptional regulator MatA	146	10	Regulation of adherence to epithelium or biofilm formation.[98]
mazE	Antitoxin MazE	146	13	Toxin/antitoxin - stress response.[99]
mazF	mRNA interferase MazF	146	13	Toxin/antitoxin - stress response.[99]
merA	Mercuric reductase	146	4	Mercury resistance.[100]
mhpC	2-hydroxy-6-oxononadienedioate/2-hydroxy-6-oxononatrienedioate hydrolase	144	0	*Mhp* cluster- recycling of carbon sources in the ecosystem.[101]
mhpD	2-keto-4-pentenoate hydratase	143	0	*Mhp* cluster- recycling of carbon sources in the ecosystem.[101]
mhpF	Acetaldehyde dehydrogenase	144	0	*Mhp* cluster- recycling of carbon sources in the ecosystem.[101]
mkaC	Virulence genes transcriptional activator	146	18	Activation of virulence genes in *Salmonella serovar Typhimurium*.[102]
mocA	Molybdenum cofactor cytidylyltransferase	145	13	Molybdopterin Cytosine Dinucleotide biosynthesis.[103]
mshB	1D-myo-inositol 2-acetamido-2-deoxy-alpha-D-glucopyranoside deacetylase	144	0	Production of mycothiol - only found in *Actinobacteria.*[104]
mtaD	5-methylthioadenosine/S-adenosylhomocysteine deaminase	145	0	-
narV	Respiratory nitrate reductase 2 gamma chain	144	0	Nitroduction cycle in the environment.[105]
narW	putative nitrate reductase molybdenum cofactor assembly chaperone NarW	145	0	Nitroduction cycle in the environment.[105]
nhoA	N-hydroxyarylamine O-acetyltransferase	143	0	-
outO	Type 4 prepilin-like proteins leader peptide-processing enzyme	19	133	-
pac	Penicillin G acylase precursor	145	20	Production of semisynthetic penicillin.[106]
pbpX	Putative penicillin-binding protein PbpX	146	13	Endospore formation and vegetative state in *Bacillus* - environmental stress response.[107]
pcaK	4-hydroxybenzoate transporter PcaK	144	0	Metabolism of aromatic compounds in Pseudomonas.[108]
pduA	Propanediol utilization protein PduA	146	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]

pduB	Propanediol utilization protein PduB	146	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]
pduC	Propanediol dehydratase large subunit	146	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]
pduD	Propanediol dehydratase medium subunit	145	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]
pduE	Propanediol dehydratase small subunit	146	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]
pduF	Propanediol diffusion facilitator	146	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]
pduL	Phosphate propanoyltransferase	146	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]
pduU	Propanediol utilization protein PduU	146	0	1,2 propanediol utilization - propylene glycol which is used in many manufactured products and shown to be a pollutant in waterways.[109,110]
pemK	mRNA interferase PemK	1	135	Stable inheritance of R100 resistance plasmid - mercuric ion sensitivity.[111]
pimB	GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol monomannoside mannosyltransferase	0	133	Cell wall synthesis - *M. tuberculosis* - can increase rate of human macrophage death.[112]
pptA	Tautomerase PptA	145	0	-
pstS1	Phosphate-binding protein PstS 1 precursor	146	12	Involved in phosphate uptake and virulence in *M. tuberculosis*.[113]
pucA	putative xanthine dehydrogenase subunit A	146	13	-
putA	Bifunctional protein PutA	146	13	Proline utilization - critical for survival of *S. aureus* in low proline environments with the host.[114]
putP	Sodium/proline symporter	145	13	Proline utilization - critical for survival of *S. aureus* in low proline environments with the host.[114]
racX	putative amino-acid racemase	0	134	-
relE2	Toxin RelE2	146	4	Toxin/antoxin operon.
rfaL	O-antigen ligase	146	13	Outer membrane biosynthesis.
rfbC	dTDP-4-dehydrorhamnose 3,5-epimerase	0	134	O antigen synthesis.[115]
rfbD	dTDP-4-dehydrorhamnose reductase	0	133	O antigen synthesis.[115]
rhmD	L-rhamnonate dehydratase	146	13	Fructose and mannose metabolism, and transcription.

rhmR	putative HTH-type transcriptional regulator RhmR	145	13	Fructose and mannose metabolism, and transcription.
rhsA	putative deoxyribonuclease RhsA	146	1	Intracellular competition.[116]
rmlA1	Glucose-1-phosphate thymidylyltransferase 1	0	134	-
rpiB	Ribose-5-phosphate isomerase B	146	13	Sugar phosphate isomerase and metabolism of rare sugar allose.[117]
salL	Adenosyl-chloride synthase	146	13	Chlorine incorporation in *Salinispora tropica* - a marine bacterial species.[118]
sbmC	DNA gyrase inhibitor	146	0	Up-regulated in *Salmonella enterica var. Typhimurium* in sub-MIC fluroquinolones.[119]
sfaG	S-fimbrial protein subunit SfaG precursor	0	131	Virulence factor.[120]
sfaS	S-fimbrial adhesin protein SfaS precursor	0	134	Virulence factor.[120]
sicA	Chaperone protein SicA	146	0	Linked to virulence in *Salmonella spp*.[121]
sinR	HTH-type transcriptional regulator SinR	145	8	Involved in biofilm formation and lactate utilization in *Bacillus subtilis.*[122]
taqIM	Modification methylase TaqI	146	0	-
torI	Response regulator inhibitor for tor operon	146	0	Regulation of environmental response torCAD opern.[123]
wcaJ	UDP-glucose:undecaprenyl-phosphate glucose-1-phosphate transferase	0	132	Lipid-linked glycan biosynthesis.[124]
wzxC	Lipopolysaccharide biosynthesis protein WzxC	0	134	-
xylF	D-xylose-binding periplasmic protein precursor	0	135	Xylose sugar transport.[125]
xylG	2-hydroxymuconic semialdehyde dehydrogenase	1	134	Xylose sugar transport.[125]
xylH	Xylose transport system permease protein XylH	0	132	Xylose sugar transport.[125]
xylR	Xylose operon regulatory protein	0	134	Xylose sugar transport.[125]
yafQ	mRNA interferase YafQ	146	0	With dinJ forms dinJ-yafQ operon - toxin/antitoxin stress response.[83]
yagU	Inner membrane protein YagU	146	13	Extreme acid tolerance.[126]
yahK	Aldehyde reductase YahK	145	0	Glucose metabolism and the production of aldehydes.[127]

ybbY	Putative purine permease YbbY	146	14	-
ybhA	Pyridoxal phosphate phosphatase YbhA	0	135	Sugar metabolism.
ydcR	putative HTH-type transcriptional regulator YdcR	146	0	-
yddE	putative isomerase YddE	145	0	-
yedR	Inner membrane protein YedR	146	13	-
ygbF	CRISPR-associated endoribonuclease Cas2	146	13	-
ygbT	CRISPR-associated endonuclease Cas1	146	13	-
ygeX	Diaminopropionate ammonia-lyase	146	13	Growth on DL-DAP as carbon source.[128]
yggF	Fructose-1,6-bisphosphatase 2 class 2	0	134	Sugar metabolism.
yhaI	Inner membrane protein YhaI	145	13	-
yjdF	Inner membrane protein YjdF	0	132	-
yjdL	putative dipeptide and tripeptide permease YjdL	146	0	Nutrient transport. [129]
yjiE	HTH-type transcriptional regulator YjiE	1	133	-
yjiG	Inner membrane protein YjiG	0	132	-
yjmD	putative zinc-type alcohol dehydrogenase-like protein YjmD	146	3	Characterized in *Bacillus subtilis.*
ymfD	Bacillibactin exporter	146	13	Iron acquisition in *Bacillus subtilus.*[130]
yqiJ	Inner membrane protein YqiJ	146	20	-
ytbE	putative oxidoreductase YtbE	146	1	Aldehyde catabolism in *Bacillus subtilis.*[131]
yvqK	Cob(I)yrinic acid a,c-diamide adenosyltransferase	146	0	-

Table 3: Genes that segregated completely or in significantly higher numbers in Vietnamese isolates of *S. sonnei (n=146)* and *S. flexneri (n=135).*

Global S. sonnei and S. flexneri pan genome content

A global set of *S. sonnei* and *S. flexneri* isolates (*n* = 60 of each species) was analysed with the same measures as Vietnamese samples, with pan genome construction and functional annotation revealing species level differences not subject to localized adaptation and gene flow that could be occurring uniquely in Vietnam.

The size of the global pan genomes of each species was analogous to the Vietnamese isolates (*S. sonnei* = 5,731 genes, *S. flexneri* = 6,296 genes), albeit with a marginally larger 'cloud' genome (genes found in less than 15% of isolates; *S. sonnei* = 2,008, *S. flexneri* = 2,531), which is to be expected due to the greater diversity, and therefore more variable potential gene pool, of the global isolates. The number of genes present in at least 95% of the tested isolates (the softcore genome) was comparable in the global and Vietnamese *S. flexneri* samples (genes in ≥95% global isolates = 2,695, genes in ≥95% Vietnamese isolates = 2,462), though the number was lower in the global *S. sonnei* softcore genome than that of the Vietnamese isolates (genes in ≥95% global isolates = 2,720, genes in ≥95% Vietnamese isolates = 3,184). Again, this is likely due to the clonal history and low genetic diversity of the Vietnamese *S. sonnei* samples used in this study.

Interestingly, of the genes that segregated exclusively, or in significantly higher numbers, in one species of the Vietnamese *Shigella spp.,* 76 of the 106 genes found in *S. sonnei* and 27 of the 37 genes in *S. flexneri* were also identified at significantly different their frequency in the analysis of global isolates. Many of the genes associated with adaptation to environmental stress and metabolic differences are conserved in the global pan genomes of each species, indicating that these species level adaptive differences identified in the Vietnamese isolates are also evident at the global scale.

There were a total of 151 genes identified in significantly different numbers in either *S. sonnei* or *S. flexneri* global isolates, with 50 of these genes further to those found through the analysis of the Vietnamese samples **(Table 4).** The majority of these genes appear to be associated with inner membrane proteins and cellular transport, including three genes within the LsrABCD transporter complex. Notably though, there were also three genes found within the ethanolamine (*eut)* operon predominately in the global *S. sonnei* pan genome. This suite of genes is involved in the metabolism of the alcohol ethanolamine, a common ingredient in industrial detergents, with these genes also found in >98% of the Vietnamese samples of *S. sonnei* and around 20-25% of *S. flexneri.*

Gene name	Roary Annotation	*Shigella sonnei* presence *(n=60)*	*Shigella flexneri* presence *(n=60)*
aarA	Rhomboid protease AarA	60	0
abgT	p-aminobenzoyl-glutamate transport protein	59	0
aceK	Isocitrate dehydrogenase kinase/phosphatase	0	60
adhT	Alcohol dehydrogenase	0	59
allA	Ureidoglycolate lyase	3	60
allC	Allantoate amidohydrolase	0	60
allR	HTH-type transcriptional repressor AllR	3	60
apllR	Type-2 restriction enzyme AplI	57	0
arfA	Alternative ribosome-rescue factor A	59	6
argF	Ornithine carbamoyltransferase	59	0
atoA	Acetate CoA-transferase subunit beta	58	0
atoD	Acetate CoA-transferase subunit alpha	59	0
atoE	Short-chain fatty acids transporter	60	0
avtA	Valine--pyruvate aminotransferase	3	60
bfd	Bacterioferritin-associated ferredoxin	0	60
bfr	Bacterioferritin	0	60
bsdC	Phenolic acid decarboxylase subunit C	0	59
cadA	Lysine decarboxylase, inducible	60	0
cadC	Transcriptional activator CadC	60	0
casD	CRISPR system Cascade subunit CasD	60	0
ccmL	Carbon dioxide concentrating mechanism protein CcmL	59	0
cdhB	Caffeine dehydrogenase subunit beta	60	1
cdhR	HTH-type transcriptional regulator CdhR	0	60
cmtB	Mannitol-specific cryptic phosphotransferase enzyme IIA component	60	4
cotSA	Spore coat protein SA	0	60
ddpA	putative D,D-dipeptide-binding periplasmic protein DdpA precursor	60	0
ddrA	Diol dehydratase-reactivating factor alpha subunit	59	5

dlgD	2,3-diketo-L-gulonate reductase	3	60
dmlA	D-malate dehydrogenase [decarboxylating]	59	0
eamA	putative amino-acid metabolite efflux pump	0	60
envR	putative acrEF/envCD operon repressor	58	5
eutD	Ethanolamine utilization protein EutD	60	5
eutM	Ethanolamine utilization protein EutL	59	5
eutN	Ethanolamine utilization protein EutM precursor	60	5
flr	Flavoredoxin	58	0
frmA	S-(hydroxymethyl)glutathione dehydrogenase	58	0
frmB	S-formylglutathione hydrolase FrmB	59	0
frmR	Transcriptional repressor FrmR	58	0
gadB	Glutamate decarboxylase beta	6	60
galD	4-oxalomesaconate tautomerase	0	58
galF	UTP--glucose-1-phosphate uridylyltransferase	0	60
glcB	Malate synthase G	0	59
gno	Gluconate 5-dehydrogenase	59	0
guaD	Guanine deaminase	60	3
hcpA	Major exported protein	60	0
hilA	Transcriptional regulator HilA	59	0
hipA	Serine/threonine-protein kinase HipA	59	0
hosA	Transcriptional regulator HosA	0	59
hyuA	D-phenylhydantoinase	60	0
iadA	Isoaspartyl dipeptidase	0	59
icaA	N-glycosyltransferase	59	0
idnD	L-idonate 5-dehydrogenase (NAD(P)( ))	60	0
idnT	Gnt-II system L-idonate transporter	60	0
iraD	Anti-adapter protein IraD	0	59
lacY	Lactose permease	57	0

lsrB	Autoinducer 2-binding protein LsrB precursor	0	59
lsrD	Autoinducer 2 import system permease protein LsrD	0	59
lsrF	putative aldolase LsrF	0	59
lsrG	Autoinducer 2-degrading protein LsrG	0	59
lysU	Lysine--tRNA ligase, heat inducible	59	0
malS	Alpha-amylase precursor	3	60
matA	HTH-type transcriptional regulator MatA	58	0
mazE	Antitoxin MazE	60	0
mazF	mRNA interferase MazF	60	0
merA	Mercuric reductase	60	0
mhpC	2-hydroxy-6-oxononadienedioate/2-hydroxy-6-oxononatrienedioate hydrolase	58	0
mhpD	2-keto-4-pentenoate hydratase	59	0
mhpE	4-hydroxy-2-oxovalerate aldolase	59	0
mhpF	Acetaldehyde dehydrogenase	59	0
mkaC	Virulence genes transcriptional activator	57	1
mocA	Molybdenum cofactor cytidylyltransferase	59	1
mscL	Large-conductance mechanosensitive channel	60	6
mshB	1D-myo-inositol 2-acetamido-2-deoxy-alpha-D-glucopyranoside deacetylase	59	0
mtaD	5-methylthioadenosine/S-adenosylhomocysteine deaminase	59	1
narV	Respiratory nitrate reductase 2 gamma chain	60	0
narY	Respiratory nitrate reductase 2 beta chain	58	2
nhoA	N-hydroxyarylamine O-acetyltransferase	58	0
outO	Type 4 prepilin-like proteins leader peptide-processing enzyme	3	60
pac	Penicillin G acylase precursor	60	0
pad1	putative aromatic acid decarboxylase	0	59
pbpX	Putative penicillin-binding protein PbpX	59	0
pduA	Propanediol utilization protein PduA	59	0
pduB	Propanediol utilization protein PduB	57	0

pduC	Propanediol dehydratase large subunit	59	0
pduD	Propanediol dehydratase medium subunit	58	0
pduE	Propanediol dehydratase small subunit	58	0
pduF	Propanediol diffusion facilitator	57	0
pduL	Phosphate propanoyltransferase	58	0
pduU	Propanediol utilization protein PduU	60	0
pemK	mRNA interferase PemK	0	60
phnR	Putative transcriptional regulator of 2-aminoethylphosphonate degradation operons	0	60
pimB	GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol monomannoside mannosyltransferase	0	60
pptA	Tautomerase PptA	57	0
pstS1	Phosphate-binding protein PstS 1 precursor	60	0
pucA	putative xanthine dehydrogenase subunit A	60	0
putA	Bifunctional protein PutA	59	0
putP	Sodium/proline symporter	57	0
relB	Antitoxin RelB	0	57
rfbC	dTDP-4-dehydrorhamnose 3,5-epimerase	0	60
rfbD	dTDP-4-dehydrorhamnose reductase	0	60
rhsA	putative deoxyribonuclease RhsA	60	0
rmlA1	Glucose-1-phosphate thymidylyltransferase 1	0	59
rnz	Ribonuclease Z	0	60
rpiB	Ribose-5-phosphate isomerase B	57	0
rspA	30S ribosomal protein S1	0	60
salL	Adenosyl-chloride synthase	60	0
sbmC	DNA gyrase inhibitor	58	0
sinR	HTH-type transcriptional regulator SinR	57	0
symE	Toxic protein SymE	0	60
tam	Trans-aconitate 2-methyltransferase	0	60
tnaB	Low affinity tryptophan permease	3	60

torI	Response regulator inhibitor for tor operon	58	0
traA	Pilin precursor	58	8
wcaJ	UDP-glucose:undecaprenyl-phosphate glucose-1-phosphate transferase	0	60
wzxC	Lipopolysaccharide biosynthesis protein WzxC	0	60
xylF	D-xylose-binding periplasmic protein precursor	3	60
xylG	2-hydroxymuconic semialdehyde dehydrogenase	3	60
xylR	Xylose operon regulatory protein	3	60
yafO	mRNA interferase YafO	0	60
yagU	Inner membrane protein YagU	57	0
yajR	Inner membrane transport protein YajR	60	6
ybhA	Pyridoxal phosphate phosphatase YbhA	0	60
ybhI	Inner membrane protein YbhI	0	58
ydcO	Inner membrane protein YdcO	57	0
yddE	putative isomerase YddE	60	0
ydjE	Inner membrane metabolite transport protein YdjE	0	60
ydjH	putative sugar kinase YdjH	0	60
ygbF	CRISPR-associated endoribonuclease Cas2	60	0
ygbM	Putative hydroxypyruvate isomerase YgbM	59	6
ygbN	Inner membrane permease YgbN	58	6
ygbT	CRISPR-associated endonuclease Cas1	60	0
ygeX	Diaminopropionate ammonia-lyase	60	0
yggF	Fructose-1,6-bisphosphatase 2 class 2	0	59
yhaI	Inner membrane protein YhaI	58	2
yhaV	Toxin YhaV	58	0
yiaO	2,3-diketo-L-gulonate-binding periplasmic protein YiaO precursor	60	1
yidK	putative symporter YidK	0	60
yigG	Inner membrane protein YigG	0	60
yjdF	Inner membrane protein YjdF	1	60

yjdL	putative dipeptide and tripeptide permease YjdL	57	0
yjiE	HTH-type transcriptional regulator YjiE	2	59
yjiG	Inner membrane protein YjiG	0	59
yjmD	putative zinc-type alcohol dehydrogenase-like protein YjmD	59	0
ykgB	Inner membrane protein YkgB	58	0
ymfD	Bacillibactin exporter	59	0
ymgF	Inner membrane protein YmgF	0	60
ynfE	Putative dimethyl sulfoxide reductase chain YnfE precursor	8	58
yoeB	Toxin YoeB	60	5
ytbE	putative oxidoreductase YtbE	57	0
yvqK	Cob(I)yrinic acid a,c-diamide adenosyltransferase	58	0
zntR	HTH-type transcriptional regulator ZntR	60	6

Table 4: Genes that segregated completely or in significantly higher numbers in global isolates of *S. sonnei* and *S. flexneri (n=60).*

5.5 Discussion

In this study, I identified differences in the presence of particular genes, and thus potential functional differences, within the individual genomes of isolates of *S. sonnei* and *S. flexneri* in Vietnam to infer population-level variation between phenotypic traits of each species, which may contribute to their niche adaptation or environmental survival (**Table 3**). Building a representative pan genome for these species that have undergone a recent shift in prevalence in this region has resulted in the identification of potential key genomic determinants of ecology and lifestyle that may, in part, explain the epidemiological differences observed in Vietnamese *Shigella*.

Many of the species-level gene differences described between the Vietnamese isolates are also found in a global subset of *S. sonnei* and *S. flexneri* strains (**Table 4**). This suggests rather than local adaptation in Vietnamese *S. sonnei* leading to a spread of the species in this region, a change in the external or host environment in Vietnam may be responsible for the recent observed patterns of species prevalence. *S. sonnei* may either be fundamentally better adapted, or faster to adapt, to this changing environment whilst *S. flexneri* infections have become rarer as they are less suited to the new conditions. This theory would be congruent with observed and anecdotal evidence of both host changes, including diet[132] and increased antibiotic use[133], and increased awareness of sanitary practices in this region[134].

An absence of many *E. coli* catabolic genes has been described in *Shigella,* including the inability to ferment lactose and mucate and the lack of decarboxylation of lysine[46,135]. The reason for these functional losses was attributed to the specialization of *Shigella* to intracellular life within the epithelia and so the breadth of catabolic pathways present in the *E. coli,* which can live freely in the intestine, is not required. Additionally, there is evidence of the antagonistic role of the *cadA* gene, involved in lysine decarboxylation, in virulence, as demonstrated through the introduction of the gene into *S. flexneri*[136]. Interestingly, the *cadA* gene (and the transcriptional regulator *cadC*) was identified here in Vietnamese *S. sonnei* along with a number of other *E. coli* catabolic genes such as those responsible for malate, idonate, lactose, and fructose metabolism (*dmlA, idnD, lacY* and *rhmD* respectively). Evidence of differential stability of the major virulence plasmid, pINV, at environmental temperatures in *S. sonnei* and *S. flexneri,* suggesting further host adaptation in *S. sonnei*[47], may also explain differences in catabolic genes present between species. It may be beneficial for *S. sonnei* to maintain a wider range of catabolic genes for persisitence in a host if the organisms is only predominately transmitted through person-to-person contact with an absence of any environmental lifestyle component.

This increased range of catabolic genes may indicate that *S. sonnei* has the ability to occupy a niche more similar to commensal *E. coli* than *S. flexneri* and is not restricted solely to intracellular survival within epithelial cells. As *Shigella spp.* are non-motile and lack a flagellum, it is possible that the presence of genes involved in biofilm formation identified here in *S. sonnei*; *matA*[98]*, hipAB*[94] and *sinR*[122] as well as the potential role of toxin/antitoxin systems, play an important part in allowing the species to live outside epithelial cells inside a host. Biofilm formation may also compensate for a loss of virulence that can occur from the antagonistic effect of catabolic genes, with evidence of increase virulence gene expression in pathogenic bacterial strains that form biofilms[137,138].

In contrast to previous suggestions that *S. sonnei* has limited persistence outside a host[47], the capacity for *S. sonnei* to form biofilms may be important for survival outside of the host cells. *Shigella spp.* are known to persist for a short time in a variety of external environments including foodstuffs[37,39,40], fomites[41] and waterways[24,29,42]. Along with the ability of biofilms to protect bacteria and boost growth[139], there is evidence of genes in the Vietnamese *S. sonnei* pan genome that could also increase survival in an external environment. These include genes associated with persistence in aquatic environments[77] (*ccmL)*, acclimation to UV-B and low temperature[79] (*clpP1)*, degradation of the herbicide atrazine[74] (*atzC),* detoxification of formaldehyde[86] (*frm* genes), and utilization of propylene glycol, a common water pollutant[109,110] (*pdu* genes). These mechanisms for adaptation to prolonged environmental survival in *S. sonnei* appears to be, in part, through gene transfer from other exogenous bacteria, which is supported by the presence of multiple genes identified in these strains with origins in *Bacillus subtilis*[107,116,122,130,131]*,* a common soil bacteria, as well as single genes from the marine bacterium *Salinispora tropica*[118] and species of aquatic cyanobacteria *Synechocystis*[77] (**Table 3).**

The potential similarity in lifestyles of *S. sonnei* and *E. coli* may indicate that factors that will determine the distribution and occurrence of commensal *E. coli,* including diet[140], may also affect the spread of *S. sonnei*. Changes to food consumption have been described in the developing world, with a move towards a higher calorie and fat and protein-rich diet[132]. Data are available for national averages of calories, fat, protein and carbohydrates consumed taken from the Food and Agriculture Organization of the United Nations[141]. Using data from Thompson *et al,* 2015 looking at the ratio between the frequency of *S. sonnei* and *S. flexneri* from 1990-2014 in 87 countries globally[142], we can estimate the predictive effect of each dietary component on the proportion of infections by each species over this time (**Figure 4**).

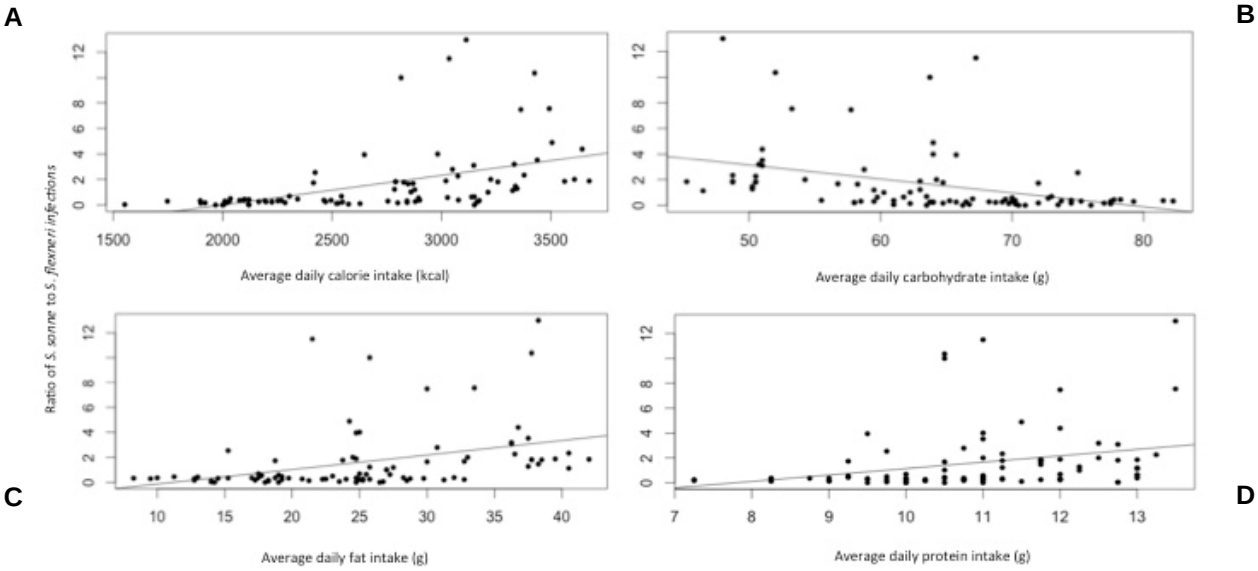


Figure 4: Comparison of the components of daily average dietary intake per person to the ratio of *S. sonnei* to *S. flexneri* infections in 87 countries. The lines represent the linear regression model for each dietary factor**. A)** Average calorie intake (kcal) (R^2 = 0.2189, p = 4.889x10^{-6}), **B)** Average carbohydrate intake (g) (R^2 = 0.1506, p = 2.039x10^{-5}), **C)** Average fat intake (g) (R^2 = 0.1416, p = 3.275x10^{-5}), and **D)** Average protein intake (g) (R^2 = 0.07676, p = 9.378x10^{-4}).

There is a significant correlation between the ratio of *S. sonnei* to *S. flexneri* infections and the combined effect of the country-level average dietary intake of fat, carbohydrates, protein and calories (R^2 = 0.2303, p < 0.001). The greatest individual contribution to this linear regression model was by the average per country calorie intake, which was positively associated with likelihood of infection by *S. sonnei* (p = 0.013) when also controlling for the other dietary factors. There is also a strong association between calorific intake and *Shigella* infection ratio when controlling for the country-level GDP[143] (p = 0.055), an indicator of socio-economic status, though there is no significant correlation between GDP and the ratio of *Shigella* infection (p = 0.542). Although this does not directly indicate that these dietary factors are having a causative effect on the frequency of infections from either *Shigella* species, the evidence for this relationship coupled with the variation in the metabolic genes between species does suggest that diet may be one factor in determining the presence of particular *Shigella* species. It has been known that diet and nutritional status can influence the composition of microbial gut flora[144] so it may be that higher calorie diets introduce a wider variety of catabolic substrates into the host gut, which *S. sonnei* is able to effectively utilize to increase growth and survival.

In addition to genes associated with metabolism and host adaptation, another important suite of genes identified in the *S. sonnei* pan genome was those involved in the CRISPR/Cas system. This system acts as an acquired immune system for prokaryotes against phages and other mobile genetic elements by recognizing and cutting sections of foreign DNA that has been incorporated into the prokaryote genome[76]. The presence of this system in *S. sonnei* could provide a defence against undesirable foreign genetic elements that can reduce the fitness of the organism.

The method of pan genome construction employed in this study offers an effective framework for investigating gene presence differences between populations of the same or closely related species, particularly in bacteria that can have highly variable genomes through recombination and transformation. The genomic differences between *S. sonnei* and *S. flexneri* identified here likely do not explain entirely the causes for the shift in species prevalence observed in Vietnam and many developing countries, though logical inferences have been made based of potential functional differences that could influence species-wide epidemiological patterns. Clearly, there is evidence of some key lifestyle differences between *Shigella* species and it is possible that a combination of genomic variation, along with the widespread use of antibiotics and potential association with other organisms, has contributed to the increased incidence of *S. sonnei* infections in this region.

Additional acknowledgements

I would like to thank Liam Shaw and Hao Weng Wu for their assistance with the genetic diversity analysis and collating of the global *Shigella* strains in this study.

References

1. Perna, N. T. *et al.* Genome sequence of enterohaemorrhagic Escherichia coli O157 : H7. *Nature* **409,** 529–533 (2001).

2. Kaas, R. S., Friis, C., Ussery, D. W. & Aarestrup, F. M. Estimating variation within the genes and inferring the phylogeny of 186 sequenced diverse Escherichia coli genomes. *BMC Genomics* **13,** 577 (2012).

3. Johnson, J. R. *et al.* Abrupt emergence of a single dominant multidrug-resistant strain of Escherichia coli. *J. Infect. Dis.* **207,** 919–28 (2013).

4. Touchon, M. *et al.* Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. *PLoS Genet.* **5,** e1000344 (2009).

5. Holden, M. T. G. *et al.* A genomic portrait of the emerfences, evolution, and global spread of methicillin-resistant Staphylococcus aureus. *Genome Res.* **23,** 653–664 (2013).

6. Okoro, C. K. *et al.* Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa. *Nat. Genet.* **44,** 1215–21 (2012).

7. Casali, N. *et al.* Evolution and transmission of drug-resistant tuberculosis in a Russian population. *Nat. Genet.* (2014). doi:10.1038/ng.2878

8. Eldholm, V. *et al.* Four decades of transmission of a multidrug-resistant Mycobacterium tuberculosis outbreak strain. *Nat. Commun.* **6,** 7119 (2015).

9. Baker, M. De novo genome assembly: what every biologist should know. *Nat. Methods* **9,** 333–337 (2012).

10. Chaisson, M. J., Pevzner, P. A., Chaisson, M. J. & Pevzner, P. A. Short read fragment assembly of bacterial genomes Short read fragment assembly of bacterial genomes. 324–330 (2008). doi:10.1101/gr.7088808

11. Pallen, M. J. & Wren, B. W. Bacterial pathogenomics. *Nature* **449,** 835–42 (2007).

12. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11,** 472–477 (2008).

13. Compeau, P. E. C., Pevzner, P. A. & Tesler, G. How to apply de Bruijn graphs to genome assembly. *Nat. Biotechnol.* **29,** 987–991 (2011).

14. Yandell, M. & Ence, D. A beginner's guide to eukaryotic genome annotation. *Nat Rev Genet* **13,** 329–342 (2012).

15. Medini, D., Donati, C., Tettelin, H., Masignani, V. & Rappuoli, R. The microbial pan-genome. *Curr. Opin. Genet. Dev.* **15,** 589–94 (2005).

16. Méric, G. *et al.* A reference pan-genome approach to comparative bacterial genomics: identification of novel epidemiological markers in pathogenic Campylobacter. *PLoS One* **9,** e92798 (2014).

17. Kweon, O. *et al.* Comparative functional pan-genome analyses to build connections between genomic dynamics and phenotypic evolution in polycyclic aromatic hydrocarbon metabolism in the genus Mycobacterium. *BMC Evol. Biol.* **15,** 21 (2015).

18. Alcaraz, L. D. & Universitaria, C. Pan-genomics : Unmasking the gene diversity

hidden in the bacteria species . PrePrints. (2014).

19. Gordienko, E. N., Kazanov, M. D., Gelfand, M. S. & Gelfand, S. Evolution of pan-genomes of Escherichia coli, Shigella spp., and Salmonella enterica. *J. Bacteriol.* **195,** 2786–92 (2013).

20. Yang, F. *et al.* Genome dynamics and diversity of Shigella species, the etiologic agents of bacillary dysentery. *Nucleic Acids Res.* **33,** 6445–58 (2005).

21. Yang, J., Chen, L., Yu, J., Sun, L. & Jin, Q. ShiBASE: an integrated database for comparative genomics of Shigella. *Nucleic Acids Res.* **34,** D398-401 (2006).

22. Shepherd, J. G., Wang, L., Reeves, P. R. & Wang, L. E. I. 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 Comparison of O-Antigen Gene Clusters of Esche. (2000). doi:10.1128/IAI.68.10.6056-6061.2000.Updated

23. Zuo, G., Xu, Z. & Hao, B. Shigella Strains Are Not Clones of Escherichia coli but Sister Species in the Genus Escherichia. *Genomics, Proteomics Bioinforma.* **11,** 61–65 (2013).

24. Faruque, S. M. *et al.* Isolation of Shigella dysenteriae Type 1 and S . flexneri Strains from Surface Waters in Bangladesh : Comparative Molecular Analysis of Environmental Shigella Isolates versus Clinical Strains Isolation of Shigella dysenteriae Type 1 and S . flexneri Strai. *Appl. Environ. Microbiol.* **68,** 3908–3913 (2002).

25. Niyogi, S. K. & Pazhani, G. P. Multiresistant Shigella species isolated from childhood diarrhea cases in Kolkata, India. *Jpn. J. Infect. Dis.* **56,** 33–34 (2003).

26. von Seidlein, L. *et al.* A multicentre study of Shigella diarrhoea in six Asian countries: disease burden, clinical manifestations, and microbiology. *PLoS Med.* **3,** e353 (2006).

27. Kotloff, K. L. *et al.* Global burden of Shigella infections: implications for vaccine development and implementation of control strategies. *Bull. World Health Organ.* **77,** 651–66 (1999).

28. Banga Singh, K.-K., Ojha, S. C., Deris, Z. Z. & Rahman, R. A. A 9-year study of shigellosis in Northeast Malaysia: Antimicrobial susceptibility and shifting species dominance. *Z. Gesundh. Wiss.* **19,** 231–236 (2011).

29. Vinh, H. *et al.* A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation. *BMC Infect. Dis.* **9,** 204 (2009).

30. Qu, F. *et al.* Genotypes and antimicrobial profiles of Shigella sonnei isolates from diarrheal patients circulating in Beijing between 2002 and 2007. *Diagn. Microbiol. Infect. Dis.* **74,** 166–70 (2012).

31. Holt, K. E. *et al.* Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44,** 1056–9 (2012).

32. Holt, K. E. *et al.* Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17522–7 (2013).

33. DuPont, H. L., Levine, M. M., Hornick, R. B. & Formal, S. B. Inoculum size in shigellosis and implications for expected mode of transmission. *J. Infect. Dis.* **159,**

1126–8 (1989).

34. Garrett, V. *et al.* A recurring outbreak of Shigella sonnei among traditionally observant Jewish children in New York City: the risks of daycare and household transmission. *Epidemiol. Infect.* **134,** 1231–6 (2006).

35. Kim, D. R. *et al.* Geographic analysis of shigellosis in Vietnam. *Health Place* **14,** 755–67 (2008).

36. Cohen, D. *et al.* Recent trends in the epidemiology of shigellosis in Israel. *Epidemiol. Infect.* 1–12 (2014). doi:10.1017/S0950268814000260

37. Bagamboula, C. F., Uyttendaele, M. & Debevere, J. Acid tolerance of Shigella sonnei and Shigella flexneri. *J. Appl. Microbiol.* **93,** 479–86 (2002).

38. Vinh, H. *et al.* Rapid emergence of third generation cephalosporin resistant Shigella spp. in Southern Vietnam. *J. Med. Microbiol.* **58,** 281–3 (2009).

39. Islam, M. S., Hasan, M. K. & Khan, S. I. Growth and survival of Shigella flexneri in common Bangladeshi foods under various conditions of time and temperature. *Appl. Environ. Microbiol.* **59,** 652–654 (1993).

40. Taylor, B. C. & Nakamura, M. Survival of Shigellae in Food. *J. Hyg. (Lond).* **62,** 303–11 (1964).

41. Colwell, R. R. Survival of S Dysenteriae Type 1 on Fomites.Pdf. **19,** 177–182 (2016).

42. Karasz, O'Reilly & Bair. © 1964 Nature Publishing Group. *Nature* **202,** 693–694 (1964).

43. Saeed, A., Abd, H., Edvinsson, B. & Sandström, G. Acanthamoeba castellanii an environmental host for Shigella dysenteriae and Shigella sonnei. *Arch. Microbiol.* **191,** 83–88 (2008).

44. Jeong, H. J. *et al.* Acanthamoeba: Could it be an environmental host of Shigella? *Exp. Parasitol.* **115,** 181–186 (2007).

45. Kubler-Kielb, J., Schneerson, R., Mocca, C. & Vinogradov, E. The elucidation of the structure of the core part of the LPS from Plesiomonas shigelloides serotype O17 expressing O-polysaccharide chain identical to the Shigella sonnei O-chain. *Carbohydr. Res.* **343,** 3123–3127 (2008).

46. Lan, R. & Reeves, P. R. Escherichia coli in disguise: Molecular origins of Shigella. *Microbes Infect.* **4,** 1125–1132 (2002).

47. McVicker, G. & Tang, C. M. Deletion of toxin–antitoxin systems in the evolution of Shigella sonnei as a host-adapted pathogen. *Nat. Microbiol.* **2,** 16204 (2016).

48. Wellcome Trust Sanger Institute. No Title. (2016). at <http://www.sanger.ac.uk/resources/downloads/bacteria/shigella.html>

49. Zerbino, D. R. Technologies. 1–13 (2011). doi:10.1002/0471250953.bi1105s31.Using

50. Bao, E., Jiang, T. & Girke, T. AlignGraph: Algorithm for secondary de novo genome assembly guided by closely related references. *Bioinformatics* **30,** 319–328 (2014).

51. Seemann, T. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* **30,** 2068–2069 (2014).

52. Page, A. J. *et al.* Roary : Rapid large-scale prokaryote pan genome analysis. 13–15 (2015).

53. R Development Core Team, R. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* **1,** 409 (2011).

54. Ondov, B. D. *et al.* Fast genome and metagenome distance estimation using MinHash. *bioRxiv* 29827 (2015). doi:10.1101/029827

55. Thomas, P. D. *et al.* PANTHER: A library of protein families and subfamilies indexed by function. *Genome Res.* **13,** 2129–2141 (2003).

56. Mi, H. *et al.* The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Res.* **33,** 284–288 (2005).

57. Rasko, D. A. *et al.* The pangenome structure of Escherichia coli: Comparative genomic analysis of E. coli commensal and pathogenic isolates. *J. Bacteriol.* **190,** 6881–6893 (2008).

58. Andrews, S. C., Harrison, P. M. & Guest, J. R. Cloning, sequencing, and mapping of the bacterioferritin gene (bfr) of Escherichia coli K-12. *J. Bacteriol.* **171,** 3940–3947 (1989).

59. Gauss, G. H. *et al.* Characterization of the Bacteroides fragilis bfr gene product identifies a bacterial DPS-like protein and suggests evolutionary links in the ferritin superfamily. *J. Bacteriol.* **194,** 15–27 (2012).

60. Sahl, J. W. *et al.* Defining the phylogenomics of Shigella species: A pathway to diagnostics. *J. Clin. Microbiol.* **53,** 951–960 (2015).

61. Leclerc, H., Edberg, S., Pierzo, V. & Delattre, J. M. Bacteriophages as indicators of enteric viruses and public health risk in groundwaters. *J. Appl. Microbiol.* **88,** 5–21 (2000).

62. Cooper, T. F. & Heinemann, J. A. Postsegregational killing does not increase plasmid stability but acts to mediate the exclusion of competing plasmids. *Proc. Natl. Acad. Sci. U. S. A.* **97,** 12643–8 (2000).

63. Wen, Y., Behiels, E. & Devreese, B. Toxin-Antitoxin systems: Their role in persistence, biofilm formation, and pathogenicity. *Pathog. Dis.* **70,** 240–249 (2014).

64. Sadeghifard, N., Soheili, S., Sekawi, Z. & Ghafourian, S. Is the mazEF toxin-antitoxin system responsible for vancomycin resistance in clinical isolates of Enterococcus faecalis? *GMS Hyg. Infect. Control* **9,** Doc05 (2014).

65. Gerdes, K. Toxin-antitoxin modules may regulate synthesis of macromolecules during nutritional stress. *J. Bacteriol.* **182,** 561–572 (2000).

66. Mochizuki, A., Yahara, K., Kobayashi, I. & Iwasa, Y. Genetic addiction: Selfish gene's strategy for symbiosis in the genome. *Genetics* **172,** 1309–1323 (2006).

67. Christensen, S. K., Mikkelsen, M., Pedersen, K. & Gerdes, K. RelE, a global inhibitor of translation, is activated during nutritional stress. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 14328–33 (2001).

68. Rather, P. N. & Orosz, E. Characterization of aarA, a pleiotrophic negative regulator of the 2'-N- acetyltransferase in Providencia stuartii. *J. Bacteriol.* **176,** 5140–5144 (1994).

69. Carter, E. L. *et al.* Escherichia coli abg genes enable uptake and cleavage of the folate catabolite p-aminobenzoyl-glutamate. *J. Bacteriol.* **189,** 3329–3334 (2007).

70. Kruger, N. & Steinbuchel, A. Identification of [i]acoR[/i], a regulatory gene for the expression of genes essential for acetoin catabolism in [i]Alcaligenes eutrophus[/i] H16. *J Bacteriol* **174,** 4391–4400 (1992).

71. Edenberg, H. J. & Ph, D. Metabolism. **30,** 5–13 (2007).

72. Ho, K. K. & Weiner, H. Isolation and Characterization of an Aldehyde Dehydrogenase Encoded by the aldB Gene of Escherichia coli Isolation and Characterization of an Aldehyde Dehydrogenase Encoded by the aldB Gene of Escherichia coli †. **187,** 1067–1073 (2005).

73. Jenkins, L. S. & Nunn, W. D. Genetic and molecular characterisation of the genes involved in short-chain fatty-acid degradation in Escherichia coli: the ato system. *J. Bacteriol.* **169,** 42–52 (1987).

74. Sajjaphan, K. *et al.* Arthrobacter aurescens TC1 atrazine catabolism genes trzN, atzB, and atzC are linked on a 160-kilobase region and are functional in Escherichia coli. *Appl. Environ. Microbiol.* **70,** 4402–4407 (2004).

75. Kanjee, U. *et al.* Linkage between the bacterial acid stress and stringent responses: the structure of the inducible lysine decarboxylase. *Embo J* **30,** 931–944 (2011).

76. Rath, D., Amlinger, L., Rath, A. & Lundgren, M. The CRISPR-Cas immune system: Biology, mechanisms and applications. *Biochimie* **117,** 119–128 (2015).

77. Badger, M. R. & Price, G. D. CO2 concentrating mechanisms in cyanobacteria: Molecular components, their diversity and evolution. *J. Exp. Bot.* **54,** 609–622 (2003).

78. Mohanty, S. K. *et al.* Delineation of the caffeine C-8 oxidation pathway in Pseudomonas sp. strain CBB1 via characterization of a new trimethyluric acid monooxygenase and genes involved in trimethyluric acid metabolism. *J. Bacteriol.* **194,** 3872–3882 (2012).

79. Porankiewicz, J., Schelin, J. & Clarke, A. K. The ATP-dependent Clp protease is essential for acclimation to UV-B and low temperature in the cyanobacterium Synechococcus. *Mol. Microbiol.* **29,** 275–283 (1998).

80. Baquero, M., Bouzon, M., Quintela, J. C., Ayala, J. A. & Moreno, F. dacD , an Escherichia coli gene encoding a novel penicillin-binding protein (PBP6b) with DD-carboxypeptidase activity . dacD , an Escherichia coli Gene Encoding a Novel Penicillin-Binding Protein (PBP6b) with DD -Carboxypeptidase Activity. **178,** 7106–7111 (1996).

81. Harris, D. R. *et al.* Preserving genome integrity: The DdrA protein of Deinococcus radiodurans R1. *PLoS Biol.* **2,** (2004).

82. Bächler, C., Schneider, P., Bähler, P., Lustig, A. & Erni, B. Escherichia coli dihydroxyacetone kinase controls gene expression by binding to transcription factor DhaR. *EMBO J.* **24,** 283–93 (2005).

83. Motiejunaite, R., Armalyte, J., Markuckas, A. & Sužiedeliene, E. Escherichia coli dinJ-yafQ genes act as a toxin-antitoxin module. *FEMS Microbiol. Lett.* **268,** 112–119 (2007).

84. Lukas, H., Reimann, J., Kim, O. Bin, Grimpo, J. & Unden, G. Regulation of aerobic

and anaerobic D-malate metabolism of Escherichia coli by the LysR-type regulator DmlR (YeaT). *J. Bacteriol.* **192,** 2503–2511 (2010).

85. Hirakawa, H. *et al.* AcrS/EnvR represses expression of the acrAB multidrug efflux genes in Escherichia coli. *J. Bacteriol.* **190,** 6276–6279 (2008).

86. Gonzalez, C. F. *et al.* Molecular basis of formaldehyde detoxification: Characterization of two S-formylglutathione hydrolases from Escherichia coli, FrmB and YeiG. *J. Biol. Chem.* **281,** 14514–14522 (2006).

87. Marolda, C. L. & Valvano, M. A. The GalF protein of Escherichia coli is not a UDP-glucose pyrophosphorylase but interacts with the GalU protein possibly to regulate cellular levels of UDP-glucose. *Mol. Microbiol.* **22,** 827–840 (1996).

88. Doty, S. L., Chang, M. & Nester, E. W. The chromosomal virulence gene, chvE, of Agrobacterium tumefaciens is regulated by a LysR family member. *J. Bacteriol.* **175,** 7880–7886 (1993).

89. Klasen, R., Bringer-meyer, S. & Sahm, H. Biochemical characterization and sequence analysis of the gluconate : NADP 5-oxidoreductase gene from Gluconobacter oxydans . Biochemical Characterization and Sequence Analysis of the Gluconate : NADP 5-Oxidoreductase Gene from Gluconobacter oxydans. **177,** 2637–2643 (1995).

90. Takamatsu, D., Bensing, B. A. & Sullam, P. M. Genes in the accessory sec locus of Streptococcus gordonii have three functionally distinct effects on the expression of the platelet-binding protein GspB. *Mol. Microbiol.* **52,** 189–203 (2004).

91. Xicohtencatl-Cortes, J. *et al.* Intestinal adherence associated with type IV pili of enterohemorrhagic Escherichia coli O157 : H7. **117,** 21–24 (2007).

92. Jørgensen, M. G., Pandey, D. P., Jaskolska, M. & Gerdes, K. HicA of Escherichia coli defines a novel family of translation-independent mRNA interferases in bacteria and archaea. *J. Bacteriol.* **191,** 1191–1199 (2009).

93. Boddicker, J. D., Knosp, B. M. & Jones, B. D. Transcription of the. *Microbiology* **185,** 525–533 (2003).

94. Hansen, S. *et al.* Regulation of the Escherichia coli HipBA toxin-antitoxin system by proteolysis. *PLoS One* **7,** (2012).

95. Bausch, C. *et al.* Sequence Analysis of the GntII (Subsidiary) System for Gluconate Metabolism Reveals a Novel Pathway for l-Idonic Acid Catabolism in Escherichia coli Sequence Analysis of the GntII (Subsidiary) System for Gluconate Metabolism Reveals a Novel Pathway fo. *J. Bacteriol.* **180,** 3704–10 (1998).

96. Merrikh, H., Ferrazzoli, A. E. & Lovett, S. T. Growth phase and (p)ppGpp control of IraD, a regulator of RpoS stability, in Escherichia coli. *J. Bacteriol.* **191,** 7436–7446 (2009).

97. Stoebel, D. M. Lack of evidence for horizontal transfer of the lac operon into Escherichia coli. *Mol. Biol. Evol.* **22,** 683–690 (2005).

98. Lehti, T. A., Bauchart, P., Dobrindt, U., Korhonen, T. K. & Westerlund-Wikström, B. The fimbriae activator MatA switches off motility in Escherichia coli by repression of the flagellar master operon flhDC. *Microbiol. (United Kingdom)* **158,** 1444–1455 (2012).

99. Amitai, S., Yassin, Y. & Engelberg-kulka, H. MazF-Mediated Cell Death in Escherichia

coli : a Point of No Return MazF-Mediated Cell Death in Escherichia coli : a Point of No Return. *J. Bacteriol.* **186,** 8295–8300 (2004).

100. Ojo, K. K. *et al.* Gram-positive merA gene in gram-negative oral and urine bacteria. *FEMS Microbiol. Lett.* **238,** 411–416 (2004).

101. Torres, B., Porras, G., Garc??a, J. L. & D??az, E. Regulation of the mhp cluster responsible for 3-(3-hydroxyphenyl)propionic acid degradation in Escherichia coli. *J. Biol. Chem.* **278,** 27575–27585 (2003).

102. Taira, S., Riikonen, P., Saarilahti, H., Sukupolvi, S. & Rhen, M. The mkaC virulence gene of the Salmonella serovar Typhimurium 96 kb plasmid encodes a transcriptional activator. *Mol. Gen. Genet. MGG* **228,** 381–384

103. Neumann, M., Mittelstädt, G., Seduk, F., Iobbi-Nivol, C. & Leimkühler, S. MocA Is a Specific Cytidylyltransferase Involved in Molybdopterin Cytosine Dinucleotide Biosynthesis in Escherichia coli. *The Journal of Biological Chemistry* **284,** 21891–21898 (2009).

104. Newton, G. L., Buchmeier, N. & Fahey, R. C. Biosynthesis and functions of mycothiol, the unique protective thiol of Actinobacteria. *Microbiol. Mol. Biol. Rev.* **72,** 471–94 (2008).

105. Moreno-vivián, C. *et al.* Prokaryotic Nitrate Reduction : Molecular Properties and Functional Distinction among Bacterial Nitrate Reductases MINIREVIEW Prokaryotic Nitrate Reduction : Molecular Properties and Functional Distinction among Bacterial Nitrate Reductases. **181,** 6573–6584 (1999).

106. Kyeong Sook Choi, Jong Ahn Kim & Hyen Sam Kang. Effects of site-directed mutations on processing and activities of penicillin G acylase from Escherichia coli ATCC 11105. *J. Bacteriol.* **174,** 6270–6276 (1992).

107. Scheffers, D. J. Dynamic localization of penicillin-binding proteins during spore development in Bacillus subtilis. *Microbiology* **151,** 999–1012 (2005).

108. Pernstich, C., Senior, L., MacInnes, K. A., Forsaith, M. & Curnow, P. Expression, purification and reconstitution of the 4-hydroxybenzoate transporter PcaK from Acinetobacter sp. ADP1(). *Protein Expression and Purification* **101,** 68–75 (2014).

109. Saxena, R. K., Anand, P., Saran, S., Isar, J. & Agarwal, L. Microbial production and applications of 1,2-propanediol. *Indian J. Microbiol.* **50,** 2–11 (2010).

110. Bobik, T. & Havemann, G. Serovar Typhimurium LT2 Includes Genes Necessary for Formation of Polyhedral Organelles Involved in Coenzyme B12-Dependent 1, 2-Propanediol Degradation. *J. …* **181,** 5967–5975 (1999).

111. Foster, T. J., Nakahara, H., Weiss, A. A. & Silver, S. Transposon A-generated mutations in the mercuric resistance genes of plasmid R100-1. *J. Bacteriol.* **140,** 167–181 (1979).

112. Torrelles, J. B. *et al.* Inactivation of Mycobacterium tuberculosis mannosyltransferase pimB reduces the cell wall lipoarabinomannan and lipomannan content and increases the rate of bacterial-induced human macrophage cell death. *Glycobiology* **19,** 743–755 (2009).

113. Peirs, P. *et al.* Mycobacterium tuberculosis with disruption in genes encoding the phosphate binding proteins PstS1 and PstS2 is deficient in phosphate uptake and demonstrates reduced in vivo virulence. *Infect. Immun.* **73,** 1898–1902 (2005).

114. Schwan, W. R. *et al.* Low-proline environments impair growth, proline transport and in vivo survival of Staphylococcus aureus strain-specific putP mutants. *Microbiology* **150,** 1055–1061 (2004).

115. Klena, J. D. & Schnaitman, C. A. Function of the rfb gene cluster and the rfe gene in the synthesis of O antigen by Shigella dysenteriae 1. *Mol. Microbiol.* **9,** 393–402 (1993).

116. Koskiniemi, S. *et al.* Rhs proteins from diverse bacteria mediate intercellular competition. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 7032–7 (2013).

117. Zhang, R. G. *et al.* The 2.2 ?? resolution structure of RpiB/AlsB from Escherichia coli illustrates a new approach to the ribose-5-phosphate isomerase reaction. *J. Mol. Biol.* **332,** 1083–1094 (2003).

118. Eustáquio, A. S., Pojer, F., Noel, J. P. & Moore, B. S. Discovery and characterization of a marine bacterial SAM-dependent chlorinase. *Nat. Chem. Biol.* **4,** 69–74 (2008).

119. Yim, G., McClure, J., Surette, M. G. & Davies, J. E. Modulation of Salmonella gene expression by subinhibitory concentrations of quinolones. *J. Antibiot. (Tokyo).* **64,** 73–78 (2011).

120. Firoozeh, F., Saffari, M., Neamati, F. & Zibaei, M. Detection of virulence genes in Escherichia coli isolated from patients with cystitis and pyelonephritis. *Int. J. Infect. Dis.* **29,** 219–222 (2014).

121. Darwin, K. H. & Miller, V. L. The putative invasion protein chaperone SicA acts together with InvF to activate the expression of Salmonella typhimurium virulence genes. *Mol. Microbiol.* **35,** 949–959 (2000).

122. Chai, Y., Kolter, R. & Losick, R. A widely conserved gene cluster required for lactate utilization in Bacillus subtilis and its involvement in biofilm formation. *J. Bacteriol.* **191,** 2423–2430 (2009).

123. Ansaldi, M., Théraulaz, L. & Méjean, V. TorI, a response regulator inhibitor of phage origin in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* **101,** 9423–8 (2004).

124. Furlong, S. E., Ford, A., Albarnez-Rodriguez, L. & Valvano, M. a. Topological analysis of the Escherichia coli WcaJ protein reveals a new conserved configuration for the polyisoprenyl-phosphate hexose-1-phosphate transferase family. *Sci. Rep.* **5,** 9178 (2015).

125. Sumiya, M., Davis, E. O., Packman, L. C., McDonald, T. P. & Henderson, P. J. Molecular genetics of a receptor protein for D-xylose, encoded by the gene xylF, in Escherichia coli. *Receptors Channels* **3,** 117–128 (1995).

126. Hayes, E. T. *et al.* Oxygen limitation modulates pH regulation of catabolism and hydrogenases, multidrug transporters, and envelope composition in Escherichia coli K-12. *BMC Microbiol.* **6,** 89 (2006).

127. Rodríguez-Verdugo, A., Gaut, B. S. & Tenaillon, O. Evolution of Escherichia coli rifampicin resistance in an antibiotic-free environment during thermal stress. *BMC Evol. Biol.* **13,** 50 (2013).

128. Kalyani, J. N., Ramachandra, N., Kachroo, A. H., Mahadevan, S. & Savithri, H. S. Functional analysis of the genes encoding diaminopropionate ammonia lyase in Escherichia coli and Salmonella enterica serovar Typhimurium. *J. Bacteriol.* **194,** 5604–5612 (2012).

129. Jensen, J. M., Ismat, F., Szakonyi, G., Rahman, M. & Mirza, O. Probing the Putative Active Site of YjdL: An Unusual Proton-Coupled Oligopeptide Transporter from E. coli. *PLoS One* **7,** (2012).

130. Dertz, E. A., Xu, J., Stintzi, A. & Raymond, K. N. Bacillibactin-mediated iron transport in Bacillus subtilis. *J. Am. Chem. Soc.* **128,** 22–23 (2006).

131. Lei, J., Zhou, Y. F., Li, L. F. & Su, X. D. Structural and biochemical analyses of YvgN and YtbE from Bacillus subtilis. *Protein Sci.* **18,** 1792–1800 (2009).

132. Schmidhuber, J. & Shetty, P. The nutrition transition to 2030. Why developing countries are likely to bear the major burden. *Food Econ. - Acta Agric. Scand. Sect. C* **2,** 150–166 (2005).

133. Nguyen, K. Van *et al.* Antibiotic use and resistance in emerging economies : a situation analysis for Viet Nam. (2013).

134. Van Minh, H., Hung, N. V., Thanh, N. H. & Yang, J. C. Assessing willingness to pay for improved sanitation in rural Vietnam. *Environ. Health Prev. Med.* **18,** 275–284 (2013).

135. Pupo, G. M., Lan, R. & Reeves, P. R. 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 (2000).

136. Day W.A., J., Fern??ndez, R. E. & Maurelli, A. T. Pathoadaptive mutations that enhance virulence: Genetic organization of the cadA regions of Shigella spp. *Infect. Immun.* **69,** 7471–7480 (2001).

137. Safadi, R. Al *et al.* Correlation between in vivo biofilm formation and virulence gene expression in Escherichia coli O104:H4. *PLoS One* **7,** (2012).

138. Naves, P. *et al.* Correlation between virulence factors and in vitro biofilm formation by Escherichia coli strains. *Microb. Pathog.* **45,** 86–91 (2008).

139. Hall-Stoodley, L., Costerton, J. W. & Stoodley, P. Bacterial biofilms: from the natural environment to infectious diseases. *Nat. Rev. Microbiol.* **2,** 95–108 (2004).

140. Tenaillon, O., Skurnik, D., Picard, B. & Denamur, E. The population genetics of commensal Escherichia coli. *Nat.Rev.Microbiol.* **8,** 207–217 (2010).

141. ChartsBin. ChartsBin Statistics collector team, 2011. *Daily Calorie Intake Per Capita* (2011).

142. Thompson, C. N., Duy, P. T. & Baker, S. The Rising Dominance of Shigella sonnei: An Intercontinental Shift in the Etiology of Bacillary Dysentery. *PLoS Negl. Trop. Dis.* **9,** e0003708 (2015).

143. International Monetary Fund. No Title. *World Economic Outlook Database, April 2016* (2016). at <http://www.imf.org/external/pubs/ft/weo/2016/01/weodata/index.aspx>

144. Hoffmann, C. *et al.* Archaea and Fungi of the Human Gut Microbiome: Correlations with Diet and Bacterial Residents. *PLoS One* **8,** (2013).

Chapter 6

General Discussion

The work presented in this thesis explores different experimental and bioinformatics approaches to examine key differences in the genetics and ecology of *Shigella sonnei* and *Shigella flexneri* in Vietnam. The underlying motivation for this work is to better understand the biological factors that have contributed to the replacement of *S. flexneri* with *S. sonnei* as the cause of bacillary dysentery, or shigellosis, in many developing countries, including Vietnam. Here, I discuss the main findings of my work in the context of this question and propose future directions of study in this area of research.

The results of the work presented here lend support to the argument that there are inherent differences in the genetics and ecology of *S. sonnei* and *S. flexneri* populations in Vietnam that may, in part, contribute to the more recent, rapid emergence of *S. sonnei* in this region. Chapter two is the first of the experimental chapters where I examined differences in *S. sonnei* and *S. flexneri* survival and resistance when exposed to varying concentrations of chlorine and SDS detergent. This conception of these experiments came from the hypothesis that a potentially higher innate resistance to these disinfectants could be influencing the persistence and spread of *S. sonnei* over *S. flexneri* in developing countries. The emergence of *S. sonnei* in these regions has coincided with increased development in many of these places[1], which in turn has seen an increase in the use of disinfectants as antimicrobials. A differential resistance to these chemicals could affect the survival and transmission of pathogens.

Individual inactivation assays were conducted on three strains of Vietnamese *S. sonnei* and *S. flexneri*, and three strains of European *S. sonnei, S. flexneri* and enteropathogenic *E. coli* to look for inherent differences in chlorine and SDS resistance. SDS did not appear to have a direct biocidal effect on any of the tested strains, with all populations able to tolerate high concentrations of the compound. Exposure to varying concentrations of the chlorine-based disinfectant calcium hypochlorite, though, resulted in different survival rates among species groups (species/location), indicating different levels of resistance to chlorination. These differences were consistent within all tested strains of a group suggesting that these resistance differences are innate in these species groups.

The highest levels of resistance were seen in *S. sonnei* strains from Vietnam, with *S. flexneri* isolates from this region showing a significantly greater susceptibility to chlorination. Additional pairwise competition assays testing each Vietnamese *S. sonnei* against Vietnamese *S. flexneri* strain also found that direct competition increased the susceptibility of *S. flexneri* to chlorination, with *S. sonnei* strains possessing a significant selective advantage. These species-level resistance differences were not seen in European strains though, which suggests that the increased resistance to these chemicals in Vietnamese *S. sonnei* strains is due to recent, local adaptation. These results are consistent with the theory that *S. sonnei* population expansion will be driven by mutations in response to strong selective pressures, such as antibiotic resistance[2,3], and thus the evolution of a greater resistance to chlorine may be contributing to the emergence of this pathogen in areas of increased disinfectant use.

Leading on from these experiments, further work was carried out to determine the mechanisms that may be employed by the bacterial cell to increase resistance to chlorination, notably the role of efflux pumps in this process. The systems have been shown to be active in the cellular stress response to toxins, including exposure to antimicrobials[4–6], and thus it is proposed that these transport proteins will also be important in determining resistance to chlorination. Inactivation assays were repeated for all Vietnamese and European *Shigella* strains in both the presence and absence of a general efflux pump inhibitor CCCP to examine the role of these systems in response to chlorine exposure. Inhibition of these pumps increased the susceptibility of strains that had previously exhibited the greatest resistance to chlorine, such as Vietnamese *S. sonnei*, but did not significantly affect strains with the greatest innate susceptibility.

These results suggest that efflux pumps are most important in these organisms in response to high concentrations of chlorine but other initial stress responses may be activated in lower levels of disinfection. This indicates that the evolution of higher levels of resistance to chlorination in *Shigella* may be determined by variation in the activity of these pumps. Analysis of mutations in the genes encoding the most common of these pump systems in the tested strains could not identify any genetic variation between the most resistant and susceptible strains that may be associated with increased resistance, though future work looking at the levels of expression in these genes may help to better understand the role of these pumps in the response to chlorine disinfection.

Chapter four of the thesis begins the first of two chapters using a large set of whole genome sequence (WGS) data of *S. sonnei* and *S. flexneri* clinical isolates from Vietnam to identify genetic variation between the species that may be contributing to the rapid spread of *S.*

sonnei in this region. This chapter used various population genetics techniques, including signals of selection through convergent evolution, to characterise antibiotic resistance mutations in species. A discriminant analysis was then carried out on antibiotic resistant and susceptible groups to identify compensatory mutations that will restore any fitness loss arising from gene changes with resistance mutations[7,8].

Differences in species level resistance was found between *Shigella* species, with *S. flexneri* harbouring a greater repertoire of variants and polymorphism in a number of other resistance conferring genes, including rifampicin resistance in *rpoB* and the regulatory gene of the multiple antibiotic resistance operon *marR*. In contrast, resistance mutations in *S. sonnei* were only detected in genes associated with fluoroquinolones, but a greater proportion of the isolates tested here showed resistance. These results were consistent with the theory that *S. sonnei* expansion can be driven by the acquisition of highly beneficial antibiotic resistance mutations and resistant strains will spread rapidly through a region, thus standing levels of variation will be relatively low in this species[3]. *S. flexneri* will persist for longer in an area with resistance having a lesser selective advantage and thus resistance tends to be acquired slower at multiple occasions in a population. It was also postulated that the evolution of compensatory mutations, particularly with fluoroquinolone resistance in *S. sonnei,* may contribute to the spread of a pathogen. There was no evidence here for compensatory mutations in Vietnamese *Shigella* populations.

Finally, WGS data for these same strains were again used to construct a representative pan-genome, comprising all genes present in a population at any frequency (core and accessory genomes)[9,10], for Vietnamese *S. sonnei* and *S. flexneri.* Species level gene content differences were then compared to look for potential functional differences between *Shigella* species that may inform the ecology and transmission of the species. This analysis identified a number of genes that were found in significantly different proportions in *S. sonnei* and *S. flexneri* populations, with many genes found only in *S. sonnei* involved in environmental niche adaptation and utilisation of a wider range of metabolic substrates. This suggests that *S. sonnei* may be better able to adapt to changing environments than *S. flexneri*, with ecological transmission of both species previously described[11–13].

The results presented in the studies contained in this thesis show species level differences between Vietnamese *S. sonnei* and *S. flexneri* in their ecology and adaptive response to selective pressures. There appears to be significant differences in chlorine resistance between *Shigella* populations in this region, with variation in efflux pump activity likely to influence the levels of resistance in these microorganisms, particularly at high concentrations of the disinfectant. In addition, genomic variation in antibiotic resistance

profiles and gene content highlights differences in the adaptive response and transmission of these species, with *S. sonnei* population expansion driven by strong selection for antimicrobial resistance. In addition, *S. sonnei* populations show a greater ability to adapt and persist in changing environments, and this may be influencing the spread of this pathogen in regions undergoing relatively rapid development.

Future work

This work explores a number of different hypotheses to better understand the reasons behind the spread of *S. sonnei* in developing regions and, while contributing some key findings on this subject, there are still clearly many possibilities of research that can be undertaken to increase our knowledge of how pathogens spread, both specifically with respect to *S. sonnei* and as a general model of bacterial transmission.

A possible direction of study would be to investigate the likely association between chlorination susceptibility and antimicrobial resistance. The experimental work presented here on chlorine resistance in *Shigella*, along with previous evidence in other bacterial taxa[14], suggests that expression and regulation of efflux pumps will influence chlorine sensitivity, with particularly high concentrations of chlorine. These pumps have been shown to play a key role in antibiotic resistance also, and thus it would be interesting to explore the effect on chlorine disinfection on antibiotic resistance acquisition.

There has been some work conducted that suggests that treatment with chlorine will induce antibiotic resistance genes in *Acinetobacter* species[14]. With the results shown here on the role of efflux pumps, as well as, differences in resistance to chlorination, work investigating the interaction between antibiotic and chlorine resistance in *Shigella* would help to understand the biological processes for this resistance. In addition, if there is an additive effect of chlorine and antibiotic resistance acquisition this may influence treatment strategies as pathogens will be coming into ever increasing contact with these compounds as we attempt to eradicate diseases through disinfectants and drugs.

Furthermore, one hypothesis put forward is that the successful emergence of *S. sonnei* in many developing countries is in part through potential environmental protection afforded to *S. sonnei* by *Acanthamoeba*[15]. This organism will phagocytise the bacterium and protect it from stressful environmental conditions. *S. flexneri,* though, has been shown to induce apoptosis when grown with this amoeba and so this species will not be protected in this way in the presence of toxins[16]. The extent to which *Acanthamoeba* will influence the resistance

of *S. sonnei* to antibiotics and disinfectants should be explored to increase our knowledge of how *S. sonnei* populations have spread. The work presented here indicates that Vietnamese *S. sonnei* has a selective advantage over *S. flexneri* in the presence of chlorine, even in the absence of the amoeba, and thus a combination of these mechanisms would help to explain the rapid expansion of *S. sonnei* in developing regions.

It would be beneficial also to conduct these experiments using *Shigella* strains obtained from another region that has been undergoing the same pattern of species replacement to determine whether these inferences of resistance differences between *S. sonnei* and *S. flexneri* are typical of these populations or down to local adaptation in Vietnam. The European strains of each species did not suggest there was the same variation in resistance between the species from this region, though testing strains from area that will be imparting similar selection pressures on *Shigella* populations would allow for better resolution of these questions.

Finally, the genomic analysis of antibiotic resistance in this thesis only considered chromosomal variants when discussing the acquisition and selective advantage of resistance. It is known that many antimicrobial resistance-determining genes are found on plasmids[17–19], as well as genes involved in virulence and other ecological characteristics. Whilst there has been work conducted investigating the plasmid profiles of *Shigella* species in India[20], it would be of interest to reconstruct and compare the plasmid genomes of the *Shigella* strains used in this thesis to better understand the complete functional and characteristic differences between *S. sonnei* and *S. flexneri* in Vietnam.

Conclusion

The recent rapid expansion of *S. sonnei* in many developing regions, and the replacement of *S. flexneri* as the primary cause of shigellosis, is a fascinating phenomenon and studying this process allows us to gain a great insight into the genetics and epidemiology of the emergence of bacterial diseases. This thesis contributes data and results to this increasingly important field of research and provides some insight into a number of key factors that may have influenced the rapid spread of this pathogen. The hope is that this work can be used to better inform management strategies for disease prevention and help to better understand the dynamics of microbial pathogen populations.

References

1. Trasande, L. *et al.* How developing nations can protect children from hazardous chemical exposures while sustaining economic growth. *Health Aff.* **30,** 2400–2409 (2011).

2. Holt, K. E. *et al.* Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proc. Natl. Acad. Sci. U. S. A.* **110,** 17522–7 (2013).

3. Connor, T. R. *et al.* Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri. *Elife* **4,** 1–16 (2015).

4. Sun, J., Deng, Z. & Yan, A. Bacterial multidrug efflux pumps: Mechanisms, physiology and pharmacological exploitations. *Biochem. Biophys. Res. Commun.* **453,** 254–267 (2014).

5. Webber, M. A. & Piddock, L. J. V. The importance of efflux pumps in bacterial antibiotic resistance. *J. Antimicrob. Chemother.* **51,** 9–11 (2003).

6. Wang, W. *et al.* High-level tetracycline resistance mediated by efflux pumps Tet(A) and Tet(A)-1 with two start codons. *J. Med. Microbiol.* **63,** 1454–1459 (2014).

7. Pantel, A. *et al.* Description of compensatory *gyrA* mutations restoring fluoroquinolone susceptibility in *Mycobacterium tuberculosis*. *J. Antimicrob. Chemother.* dkw169 (2016). doi:10.1093/jac/dkw169

8. Maisnier-Patin, S. & Andersson, D. I. Adaptation to the deleterious effects of antimicrobial drug resistance mutations by compensatory evolution. *Res. Microbiol.* **155,** 360–9 (2004).

9. Tettelin, H., Riley, D., Cattuto, C. & Medini, D. Comparative genomics: the bacterial pan-genome. *Curr. Opin. Microbiol.* **11,** 472–477 (2008).

10. Alcaraz, L. D. & Universitaria, C. Pan-genomics : Unmasking the gene diversity hidden in the bacteria species . PrePrints. (2014).

11. Faruque, S. M. *et al.* Isolation of Shigella dysenteriae Type 1 and S . flexneri Strains from Surface Waters in Bangladesh : Comparative Molecular Analysis of Environmental Shigella Isolates versus Clinical Strains Isolation of Shigella dysenteriae Type 1 and S . flexneri Strai. *Appl. Environ. Microbiol.* **68,** 3908–3913 (2002).

12. Garrett, V. *et al.* A recurring outbreak of Shigella sonnei among traditionally observant Jewish children in New York City: the risks of daycare and household transmission. *Epidemiol. Infect.* **134,** 1231–6 (2006).

13. von Seidlein, L. *et al.* A multicentre study of Shigella diarrhoea in six Asian countries: disease burden, clinical manifestations, and microbiology. *PLoS Med.* **3,** e353 (2006).

14. Prasad Karumathil, D., Yin, H. B., Kollanoor-Johny, A. & Venkitanarayanan, K. Effect of chlorine exposure on the survival and antibiotic gene expression of multidrug resistant Acinetobacter baumannii in water. *Int. J. Environ. Res. Public Health* **11,** 1844–1854 (2014).

15. Saeed, A., Abd, H., Edvinsson, B. & Sandström, G. Acanthamoeba castellanii an

environmental host for Shigella dysenteriae and Shigella sonnei. *Arch. Microbiol.* **191,** 83–88 (2008).

16. Saeed, A., Johansson, D., Sandström, G. & Abd, H. Temperature Depended Role of Shigella flexneri Invasion Plasmid on the Interaction with Acanthamoeba castellanii. *Int. J. Microbiol.* **2012,** 917031 (2012).

17. Holt, K. E. *et al.* Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe. *Nat. Genet.* **44,** 1056–9 (2012).

18. Venkatesan, M. M. *et al.* Complete DNA Sequence and Analysis of the Large Virulence Plasmid of Shigella flexneri. **69,** 3271–3285 (2001).

19. Toro, C. S. *et al.* Genetic analysis of antibiotic-resistance determinants in multidrug-resistant Shigella strains isolated from Chilean children. *Epidemiol. Infect.* **133,** 81–86 (2005).

20. Dutta, S. *et al.* Shifting serotypes, plasmid profile analysis and antimicrobial resistance pattern of shigellae strains isolated from Kolkata, India during 1995-2000. *Epidemiol. Infect.* **129,** 235–43 (2002).

Supplementary Material – Chapter 2

By strain average (triplicate)															
Disinfectant concentration	**0mg/l $Ca(ClO)_2$**			**2.5mg/l $Ca(ClO)_2$**			**5mg/l $Ca(ClO)_2$**			**10mg/l $Ca(ClO)_2$**			**20mg/l $Ca(ClO)_2$**		
Time	**5mins**	**15mins**	**30mins**	**5mins**	**15mins**	**30mins**	**5mins**	**15mins**	**30mins**	**5mins**	**15mins**	**30mins**	**5mins**	**15mins**	**30mins**
S.f1266	1.6	1.02	2.93999	1.32	1.26	0.86	0.24	0.014	0.0106	0.00096	0.00018	0.00000	0.00016	6.9E-05	0.00000
S.f262-78	0.25352	0.82817	0.89577	0.34648	0.43944	0.53239	0.02423	0.01014	0.00363	0.00012	2.25352 E-06	2.8169E -07	1.69014 E-06	1.52113 E-06	1.69014 E-06
S.f9-63	1.43925	1.26168	1.18692	1.56075	1.03738	1.03738	1.29907	0.11308	0.11308	0.00042	2.42991 E-05	2.42991 E-05	0.00093 4579	3.17757 E-05	3.17757 E-05
S.f-MS	1.03571	3.21429	1.20238	1.20238	1.14286	1.20238	0.79762	0.15476	0.83333	0.05595	0.00011	0.00000	0.00000	0.00000	0.00000
S.f-DE	0.51000	0.84000	0.66500	0.50500	0.22500	0.17500	0.30000	0.20000	0.04500	0.00027	0.00078	0.00000	3.5E-06	0.00000	0.00000
S.f-EG	1.35849	0.62893	0.43396	0.72327	0.60377	0.58491	0.80503	0.37735	0.86792	0.11006	0.00040	6.28931 E-07	2.51572 E-06	0.00000	0.00000
S.s54210	0.23864	0.88636	0.52652	0.67803	0.48485	0.38258	0.00545	0.00341	0.00341	0.00017	8.52273 E-06	0.00000	1.13636 E-06	3.78788 E-07	0.00000
S.s88-83	1.52778	2.36111	1.83333	3.40278	3.16667	2.33333	0.03000	0.13472	0.04306	0.00013	0.00000	0.00000	8.33333 E-06	0.00000	0.00000
S.s43-74	1.21154	1.00962	1.00000	1.93269	2.50962	0.92308	0.03000	0.19519	0.01154	2.30769 E-05	6.73077 E-06	0.00000	0.00000	0.00000	0.00000
S.sMS	0.90426	1.38298	1.08511	0.75887	0.85106	1.11348	0.73050	0.06383	0.43972	0.02777	0.00298	0.00019	0.904255 319	2.2695E-05	2.12766E -05
S.sDE	1.35484	0.53917	0.61751	1.11521	0.32258	0.47005	0.71429	0.38710	0.70968	0.01018	1.84332 E-06	1.84332 E-06	0.00046	0.00025	0.00034
S.sEG	0.37853	0.76271	0.96045	0.73446	0.96045	0.53672	0.59322	0.27966	0.35593	0.04633	0.03051	4.51977 E-05	2.82486 E-05	0.00051	8.47458 E-05
E.c12	1.08000	2.68000	1.24000	0.34667	2.04000	0.94667	0.69333	0.80000	0.70667	0.21733	0.05333	0.24933	0.02427	0.00200	0.00002
E.c13	0.61765	0.56618	0.75735	0.71324	0.39706	0.61765	0.91912	1.47794	0.74265	0.22206	0.23824	0.18897	0.00110	5.51471 E-05	0.00000
E.c27	0.61806	0.64583	1.02778	0.59722	0.70833	0.61806	0.71528	0.36458	0.93750	0.25069	0.93750	0.26944	0.02847	0.00035	2.08333 E-06

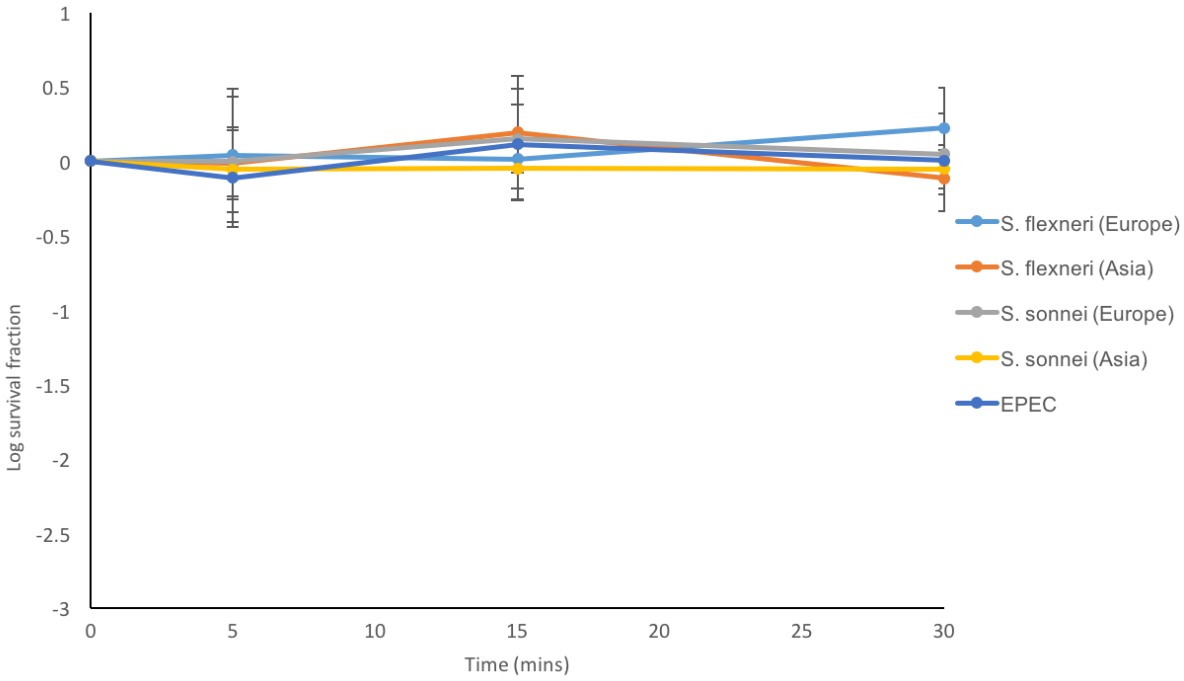
By species group average															
S. flexneri (Europe)	1.09759	1.03662	1.67423	1.07574	0.91227	0.80993	0.52110	0.04574	0.04244	0.00050	6.88509 E-05	8.19E-06	0.00037	3.40989 E-05	1.11553 E-05
S. flexneri (Asia)	0.96807	1.56107	0.76711	0.81022	0.65721	0.65410	0.63422	1.37612	0.58209	0.05543	0.00043	2.10E-07	2.00524 E-06	0.00000	0.00000
S. sonnei (Europe)	0.99265	1.41903	1.11995	2.00450	2.05371	1.21300	0.02182	0.11111	0.01933	0.00011	5.0845E -06	0.00000	3.15657 E-06	1.26263 E-07	0.00000
S. sonnei (Asia)	0.87921	0.89495	0.88769	0.86951	0.71137	0.70675	0.67933	0.24353	0.50178	0.02810	0.01116	0.00008	0.30158	0.00026	0.00015
EPEC	0.77190	1.29734	1.00838	0.55237	1.04846	0.72746	0.77591	0.88084	0.79560	0.23003	0.40969	0.23592	0.01795	0.00080	6.91667 E-06

Supplementary Table 2.1. Survival fraction from initial CFU counts at each collected time point for varying concentrations of calcium hypochlorite. Survival fractions are shown by strain and by species group.

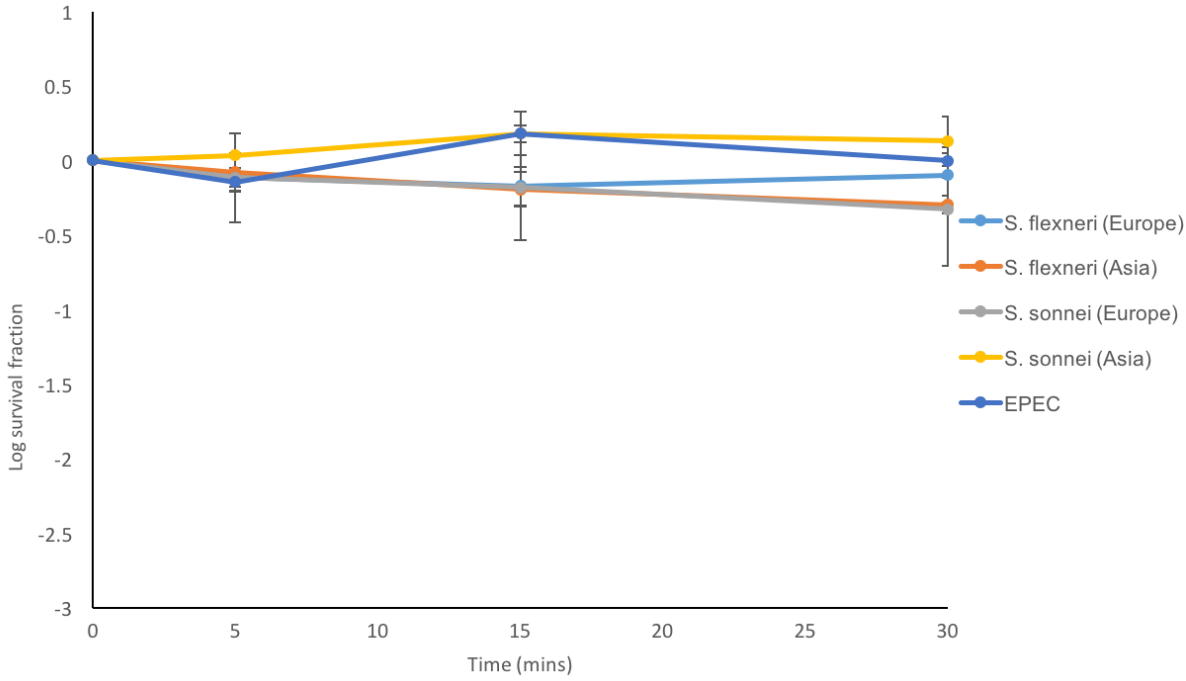
By strain average (triplicate)												
Disinfectant concentration	**0% SDS**			**15% SDS**			**22.5% SDS**			**30% SDS**		
Time	**5 mins**	**15 mins**	**30 mins**	**5 mins**	**15 mins**	**30 mins**	**5 mins**	**15 mins**	**30 mins**	**5 mins**	**15 mins**	**30 mins**
S.f1266	0.7414	0.9483	0.9425	0.0356	0.0328	0.0356	0.0053	0.0057	0.0037	0.0016	0.0015	0.0017
S.f262-78	0.9000	0.5400	1.1000	0.0287	0.2250	0.1520	0.0090	0.0102	0.0085	0.0014	0.0022	0.0019
S.f9-63	0.6782	0.5977	0.4943	0.0249	0.2333	0.2586	0.0062	0.0310	0.0087	0.0005	0.0007	0.0005
S.f-MS	0.6094	0.8672	0.5781	0.0469	0.0664	0.0727	0.0508	0.0914	0.0922	0.0084	0.0115	0.0060
S.f-DE	1.0970	0.5697	0.4364	0.0545	0.1164	0.0733	0.0406	0.0424	0.0291	0.0076	0.0053	0.0070
S.f-EG	0.8710	0.5269	0.5161	0.0930	0.0731	0.0737	0.0952	0.0306	0.0274	0.0029	0.0051	0.0067
S.s54210	0.6615	0.3438	0.2865	0.0691	0.0604	0.0882	0.0240	0.0192	0.0168	0.0005	0.0006	0.0011
S.s88-83	1.6286	1.6143	1.2857	0.2886	0.3086	0.1386	0.0059	0.0133	0.0114	0.0020	0.0021	0.0020
S.s43-74	0.4222	0.5111	0.2815	0.1200	0.1467	0.1037	0.0024	0.0187	0.0158	0.0002	0.0017	0.0013
S.sMS	0.7333	1.4571	0.8952	0.1810	0.1038	0.1867	0.1552	0.1229	0.1562	0.0695	0.0829	0.0317
S.sDE	1.3684	1.1053	1.4561	0.2395	0.1912	0.2509	0.1553	0.0324	0.1623	0.0991	0.0776	0.0808
S.sEG	1.2667	2.1600	1.8933	0.2213	0.1880	0.2347	0.1947	0.1740	0.1460	0.0229	0.0387	0.0633
E.c12	0.7500	1.3158	1.1316	0.1461	0.1500	0.1829	0.1303	0.2132	0.2355	0.1408	0.1618	0.0803
E.c13	0.7905	1.5714	1.1905	0.1114	0.1257	0.2286	0.0590	0.0924	0.0914	0.1276	0.0667	0.1029
E.c27	0.6167	1.6750	0.7333	0.0767	0.1133	0.0908	0.1800	0.0825	0.0325	0.0558	0.1083	0.2092
By species group average												
S. flexneri (Europe)	0.7732	0.6953	0.8456	0.0298	0.1637	0.1488	0.0069	0.0156	0.0070	0.0012	0.0015	0.0014
S. flexneri (Asia)	0.8591	0.6546	0.5102	0.0648	0.0853	0.0732	0.0622	0.0548	0.0496	0.0063	0.0073	0.0066
S. sonnei (Europe)	0.9041	0.8230	0.6179	0.1592	0.1719	0.1102	0.0108	0.0171	0.0147	0.0009	0.0015	0.0015
S. sonnei (Asia)	1.1228	1.5741	1.4149	0.2139	0.1610	0.2241	0.1684	0.1097	0.1548	0.0638	0.0664	0.0586
EPEC	0.7190	1.5207	1.0185	0.1114	0.1297	0.1674	0.1231	0.1293	0.1198	0.1081	0.1123	0.1308

Supplementary Table 2.2. Survival fraction from initial CFU counts at each collected time point for varying concentrations of sodium dodecyl sulphate (SDS). Survival fractions are shown by strain and by species group.

(A)

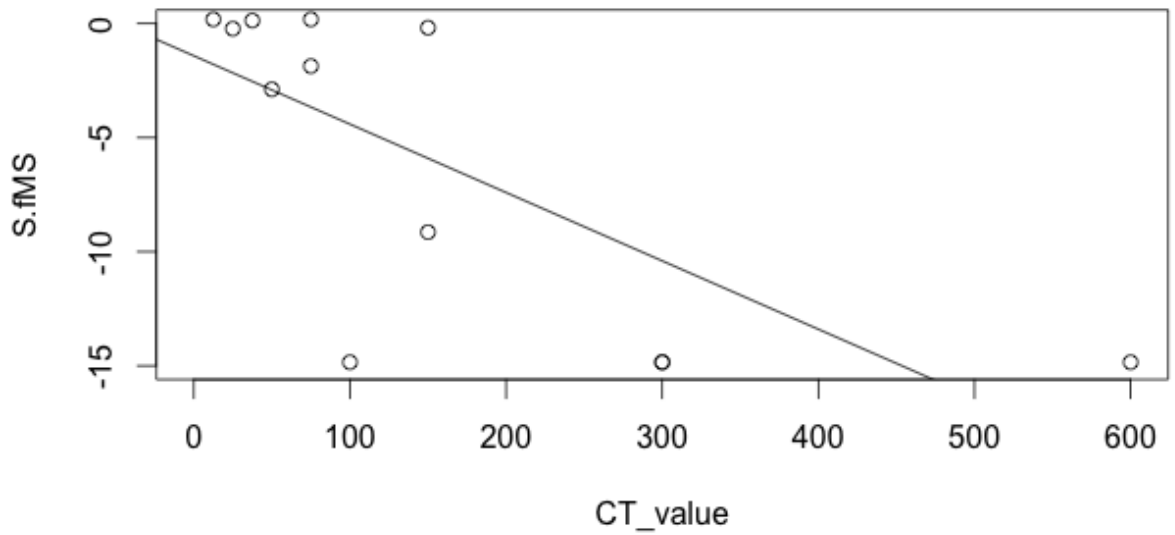


(B)

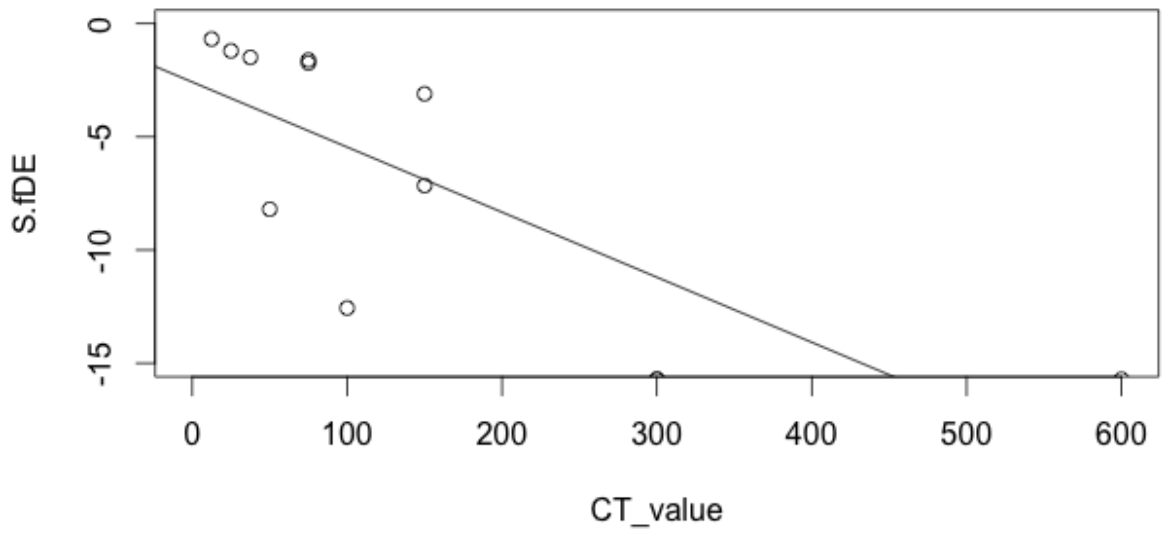


Supplementary Figure 2.1. Survival of *Shigella* species and *EPEC* against contact time in minutes of control experiments. The log survival fraction (N/N_0) each species group is shown for the control experiments when testing against (**A)** calcium hypochlorite, and (**B**) SDS. Errors bars denote the standard deviation at each recorded point.

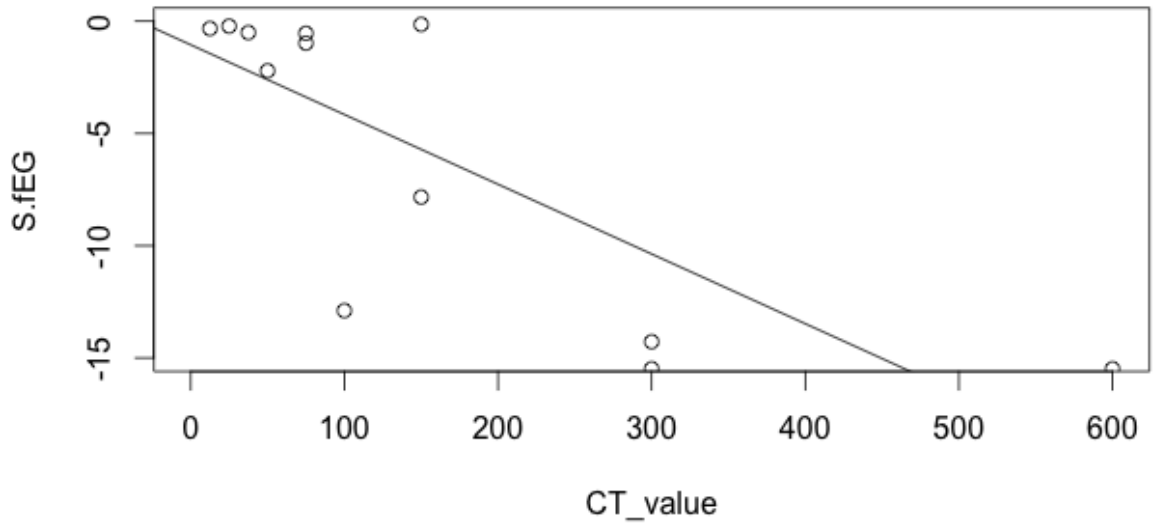
(A)


S.fMS
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

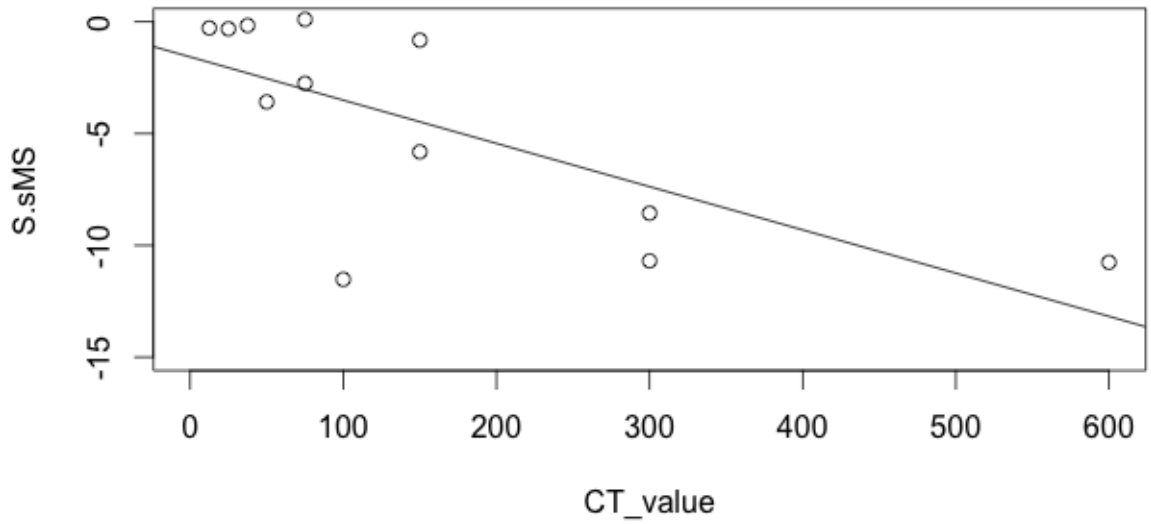
(B)


S.fDE
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

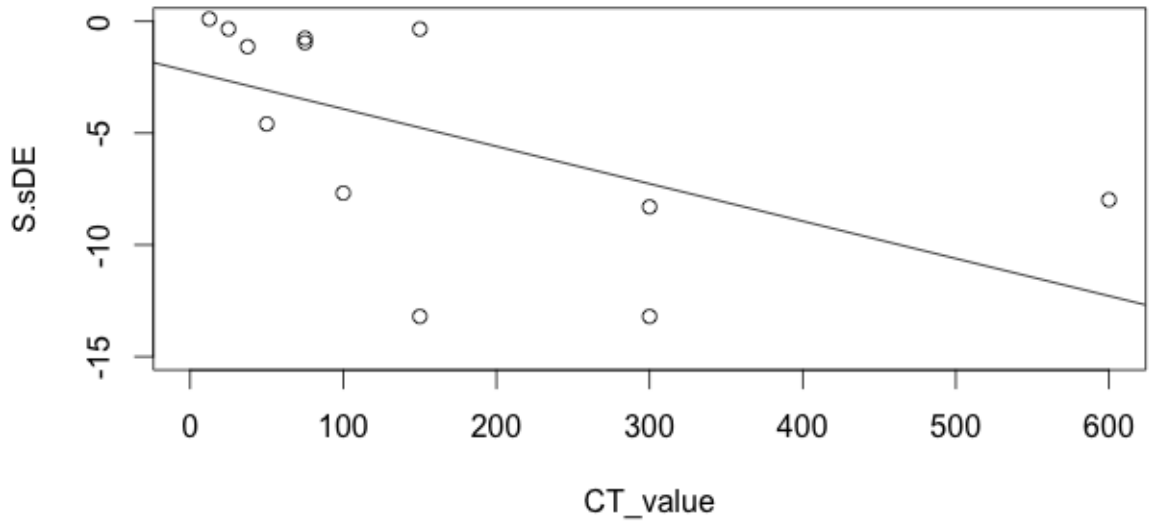
(C)


S.fEG
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

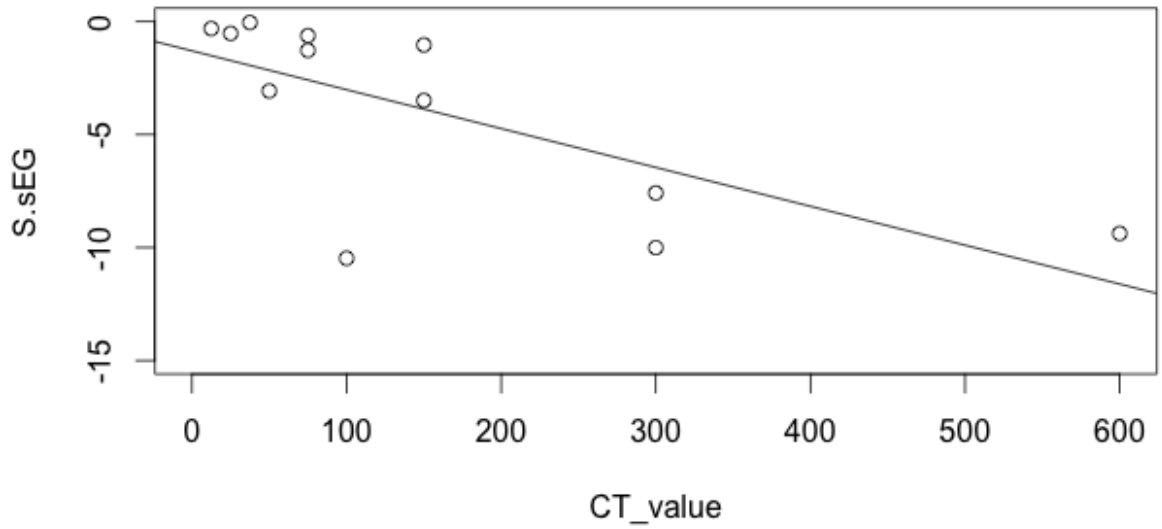
(D)


S.sMS
0
-5
-10
-15
0
100
200
300
400
500
600
CT_value

(E)



(F)



Supplementary Figure 2.2. Plots showing the natural log survival fraction (*ln (N/N$_0$)*) against CT value (mg min L $^{-1}$) for (A) *S. flexneri* MS0052, (B) *S. flexneri* DE0350, (C) *S. flexneri* EG419, (D) *S. sonnei* MS004, (E) *S. sonnei* DE1208, and (F) *S. sonnei* EG0430. All plots show the fitted least squares regression line.

Chlorine concentration		5mg/l			2.5mg/l			1.25mg/l		
Time (hours)		0.5	4	24	0.5	4	24	0.5	4	24
S.f-MS/S.S.sMS	S.sMS	0.6042	5.0000	10033.333	1.3214	8.8571	28380.952	6.3158	23.8596	38245.614
	S.f-MS	0.0180	0.0062	0.6462	0.1838	0.2059	196.0784	1.1571	4.5714	10434.286
S.f-MS/S.S.sDE	S.sDE	0.2347	3.7373	5480.0000	0.8828	15.4688	27187.500	3.8810	16.5079	39682.539
	S.f-MS	0.0001	0.0000	0.0002	0.0502	0.4010	36.0000	1.2794	2.7059	2705.8824
S.f-MS/S.S.sEG	S.sEG	2.8762	4.7525	5940.5941	3.6711	10.2645	16842.105	5.3750	18.5000	58333.333
	S.f-MS	0.1750	0.0625	0.0006	0.0969	0.3108	155.3846	2.4667	9.5844	43333.333
S.f-DE/S.S.sMS	S.sMS	2.8072	3.9959	9484.5361	3.0612	10.7143	51938.775	11.3861	14.7525	13762.376
	S.f-DE	0.0418	0.0021	6.1225	0.1472	0.4222	18.1111	1.1455	1.6545	705.4545
S.f-DE/S.S.sDE	S.sDE	1.8668	2.7356	10344.828	3.2651	4.8313	6385.5422	5.5052	14.7423	10412.371
	S.f-DE	0.0019	0.0000	0.0000	0.3405	0.0153	7.2222	0.6672	0.1463	217.9104
S.f-DE/S.S.sEG	S.sEG	0.9400	1.5859	8471.6187	0.1216	2.5992	6779.6610	5.8163	9.0816	13877.551
	S.f-DE	0.0048	0.0000	0.0000	0.0072	0.0069	0.2019	1.3953	1.4070	2351.1628
S.f-EG/S.S.sMS	S.sMS	1.2755	1.2772	20851.061	9.4167	11.9608	27156.862	6.0914	17.6344	30322.581
	S.f-EG	0.0251	0.0056	0.0152	0.2896	1.9112	2.2596	1.8605	4.5198	20120.930
S.f-EG/S.S.sDE	S.sDE	0.2959	0.9194	13775.510	2.5806	11.1236	30000.000	2.6271	24.7458	58474.576
	S.f-EG	0.0505	0.0002	0.1298	0.7466	1.9865	12801.802	2.9286	23.7743	27571.429
S.f-EG/S.S.sEG	S.sEG	0.6786	4.6429	9880.9524	3.9189	11.3514	49324.324	2.9227	15.9794	41134.021
	S.f-EG	0.0458	0.0001	3.7368	0.4237	0.3390	11016.949	1.9032	38.0323	71612.903

Supplementary Table 2.3. Survival fraction of each *Shigella* species in pairwise competition from initial CFU counts at each collected time point for varying concentrations of calcium hypochlorite.

Supplementary Materials – Chapter 3

(A)

	0mg/l			2.5mg/l			5mg/l			10mg/l		
Time (mins)	**5**	**15**	**30**	**5**	**15**	**30**	**5**	**15**	**30**	**5**	**15**	**30**
No CCCP												
S.f1266	0.639534872	1.337209302	1.569767442	1.465116279	1.593023256	0.802325581	0.073255814	1.38372093	0.006162791	0.002790698	0.000104651	8.02326E-06
S.f262-78	0.896414343	0.653386454	0.418326693	0.394422311	0.394422311	0.278884462	0.023505976	0.545816733	0.005139442	0.00025498	3.18725E-06	0
S.f9-63	0.642857143	1.032967033	1.186813187	0.835164725	1.208791209	0.626373626	0.069230769	1.483516484	0.006098901	0.001098901	2.85714E-05	4.28571E-05
S.f-MS	1.535714286	1.142857143	0.785714286	1.619047619	0.011904762	1.035714286	0.797619048	0.011428571	0.008333333	0.148809524	0.000107143	0
S.f-DE	0.831325301	1.759036145	0.704819277	0.686746988	0.746987952	1.228915663	1.192771084	0.130120482	0.043373494	0.011566265	0.001879518	3.61446E-06
S.f-EG	1.318181795	1.431818182	4.568181818	1.386363636	0.931818182	1.5	0.795454545	0.081818182	0.15	0.009090909	0.001431818	0
S.s54210	1.690909091	2.036363636	2.345454545	1.854545455	0.854545455	1.636363636	0.056181818	0.084	0.070909091	0.002036364	4.09091E-05	0
S.s88-83	1.18666664	1.32	1.786666666	1.76	1.106666667	1.333333332	0.0172	0.004	0.001773333	0.000164	0.000004	1.33333E-06
S.s43-74	1.428571357	2.071428571	4	1.714285714	0.785714286	2.714285714	0.058571429	0.037857143	0.016428571	0.000714286	1.66667E-05	7.14286E-07
S.sMS	1.149425276	1	0.827586207	0.770114943	0.977011494	1.551724138	1.183908046	0.275862069	0.67816092	0.1	0.004827586	
S.sDE	2.014925373	1.76119403	1.567164179	2.104477612	1.343283582	2.28358209	2.313432836	1.537313433	1.925373134	0.488059701	0.005970149	0.000234433
S.sEG	1.140624984	1.6875	2.640625	1.5625	0.921875	1.453125	1.640625	0.796875	0.9140625	0.15625	0.084375	0.000125

(B)

	0mg/l			2.5mg/l			5mg/l			10mg/l		
Time (mins)	**5**	**15**	**30**	**5**	**15**	**30**	**5**	**15**	**30**	**5**	**15**	**30**
						CCCP						
S.f1266	1.14893617	1.489361681	2.170212766	1.489361702	2.042553191	1.553191487	0.153191489	1.304255319	0.017085064	0.001021277	9.14894E-06	0
S.f262-78	0.999999986	2.095890411	1.780821916	0.863013699	2.123287671	0.863013699	0.036986301	1.041095 89	0.005342466	0.000123288	6.16438E-06	4.10959E-07
S.f9-63	0.924999975	0.825	1.5625	1.2875	0.8625	0.8625	0.24375	0.5625	0.0009	0.00007125	0.000000375	0
S.f-MS	0.736842105	1.473684211	1	2.842105263	0.121052632	1.754385965	0.017017544	0.001578947	0.023157895	0.016842105	0.000121053	1.75439E-06
S.f-DE	1.686274471	1.529411765	1.882352941	1.352941176	0.549019608	2.019607843	0.064705882	1.333333333	0.125490196	0.010588235	0.001372549	0
S.f-EG	1.207792208	1.051948052	0.857142857	1.636363636	0.132467532	0.935064935	0.000844156	7.4026E-05	0.003896104	0.008181818	7.79221E-06	0
S.s54210	1.075	0.525	0.625	0.7	0.369333333	0.583333333	0.023	0.000216667	0.025	0.00035	0	0
S.s88-83	0.783333333	1.966666667	0.877777777	0.955555555	0.722222222	1.433333333	0.003633333	0.000122222	0.010666667	0.000136667	0	0
S.s43-74	1.843137235	1	1.882352941	1.745098039	0.3	1.117647059	0.005882353	0.000131373	0.011176471	0.000588235	1.23529E-05	0
S.sMS	2.105263158	1.473684211	2.105263158	1.684210526	1.192982456	1.175438596	1.017543 86	0.017543 86	1.736842105	0.152631579	0.001210526	4.73684E-05
S.sDE	0.730769231	1.576923077	1.192307692	0.897435896	1.333333333	1.333333333	2.576923077	0.858974359	0.756410256	0.088461538	0.000858974	5.12821E-06
S.sEG	0.765	1.35	1	1.009999998	0.49	0.979999998	1	0.5	0.9	0.1	0.063	3.5E-05

Supplementary Table 3.1. Survival fraction of Vietnamese *Shigella* strains from initial CFU/ml counts in varying concentrations of calcium hypochlorite, (A) in the absence of the efflux pump inhibitor CCCP, and (B) in the presence of CCCP.

SNP position	Gene name	S.sEG	S.sMS	S.sDE	S.s54210	S.s88.83	S.s43.74
476077	*acrB*	0	0	0	0	1	0
2690907	*acrD*	1	0	1	0	0	0
2974428	*emrB*	0	0	0	0	0	1
3332131	*tolC*	0	0	0	0	0	1
3797699	*emrD*	0	0	0	0	0	1
3798032	*emrD*	1	0	0	0	0	1
SNP position	**Gene name**	**S.fMS**	**S.fDE**	**S.fEG**	**S.f-9.63**	**S.f-12.66**	**S.f262.78**
418085	*acrB*	1	0	0	0	0	0
418151	*acrB*	1	0	0	0	0	0
418245	*acrB*	1	0	0	0	0	0
418300	*acrB*	0	1	0	0	0	0
419126	*acrB*	0	1	0	0	0	0
419309	*acrB*	1	0	0	0	0	0
419387	*acrB*	1	1	0	0	0	0
420080	*acrB*	1	1	0	0	0	0
420725	*acrB*	1	1	1	1	1	1
420726	*acrB*	1	1	1	1	1	1
420792	*acrB*	1	1	1	1	1	1
420795	*acrB*	1	1	1	1	1	1
420796	*acrB*	1	1	1	1	1	1
420836	*acrB*	0	1	0	0	0	0
421479	*acrA*	1	1	0	0	0	0
421546	*acrA*	1	1	0	0	0	1
421726	*acrA*	1	0	0	0	0	0
421980	*acrA*	1	1	0	0	0	1
422862	*acrR*	1	0	0	0	0	0
422927	*acrR*	0	1	0	0	0	0
829423	*cmr*	0	1	0	0	0	0
829448	*cmr*	1	0	0	0	0	0
829450	*cmr*	1	1	0	0	0	0
829535	*cmr*	0	1	0	0	0	0
829685	*cmr*	1	1	0	0	0	0
830202	*cmr*	0	0	0	0	0	1
1596294	*marR*	0	0	0	0	0	1
1596505	*marR*	1	1	1	1	1	1
1723229	*ydhE*	0	0	1	0	0	0

1723527	*ydhE*	1	1	0	0	0	0
2575697	*acrD*	1	1	0	0	0	1
2576348	*acrD*	0	1	0	0	0	0
2576537	*acrD*	1	1	0	0	0	0
2577110	*acrD*	1	1	0	0	0	0
2577159	*acrD*	0	1	0	0	0	0
2577447	*acrD*	0	1	0	0	0	0
2577684	*acrD*	0	1	0	0	0	0
2577730	*acrD*	0	0	0	0	0	1
2577866	*acrD*	1	0	0	0	0	0
2578106	*acrD*	0	1	0	0	0	0
2578225	*acrD*	1	0	0	0	0	0
2578256	*acrD*	0	1	0	0	0	0
2578516	*acrD*	1	1	0	0	0	0
2786597	*emrB*	1	1	0	0	0	0
2786662	*emrB*	1	0	0	0	0	0
2786791	*emrB*	1	1	0	0	0	0
2787076	*emrB*	1	1	0	0	0	0
2787491	*emrB*	1	0	0	0	0	0
2787596	*emrB*	1	1	0	0	0	0
2787639	*emrB*	1	1	0	0	0	0
3171994	*tolC*	0	1	0	0	0	0
3172013	*tolC*	1	1	0	0	0	0
3172219	*tolC*	1	0	0	0	0	0
3172386	*tolC*	1	1	1	1	1	1
3898432	*emrD*	0	1	0	0	0	0

Supplementary Table 3.2. Full list of all SNPs found in known efflux pump genes in the experimental *S. sonnei* and *S. flexneri* strains.

Supplementary Material – Chapter 4 & 5

Species	Assension No.	Sample Name	Location	Date of isolation	Fluoroquinolone resistance (phenotypic testing)	Fluoroquinolone resistance (*in silico* - ARIBA)
S.flexneri	ERR047222	HUE49	Hue	2009	-	Resistant
S.flexneri	ERR047225	HUE54	Hue	2010	-	Susceptible
S.flexneri	ERR047231	HUE61	Hue	2010	-	Resistant
S.flexneri	ERR047233	HUE63	Hue	2010	-	Resistant
S.flexneri	ERR047235	HUE65	Hue	2010	-	Resistant
S.flexneri	ERR047236	HUE66	Hue	2010	-	Susceptible
S.flexneri	ERR047239	KH01	Khanh	2009	-	Resistant
S.flexneri	ERR047272	KH39	Khanh	2009	-	Resistant
S.flexneri	ERR047278	KH47	Khanh	2010	-	Susceptible
S.flexneri	ERR047279	KH48	Khanh	2010	-	Resistant
S.flexneri	ERR047280	KH49	Khanh	2010	-	Resistant
S.flexneri	ERR047281	KH50	Khanh	2010	-	Resistant
S.flexneri	ERR047282	KH52	Khanh	2010	-	Resistant
S.flexneri	ERR047287	KIEM	Khanh	2008	-	Resistant
S.flexneri	ERR047288	MS0001	HCMC	1995	-	Susceptible
S.flexneri	ERR047289	MS0002	HCMC	1995	-	Susceptible
S.flexneri	ERR047290	MS0005	HCMC	1995	-	Susceptible
S.flexneri	ERR047291	MS0010	HCMC	1995	-	Resistant
S.flexneri	ERR047292	MS0012	HCMC	1995	-	Susceptible
S.flexneri	ERR047293	MS0014	HCMC	1995	-	Susceptible
S.flexneri	ERR047294	MS0019	HCMC	1995	-	Susceptible
S.flexneri	ERR047295	MS0020	HCMC	1995	-	Susceptible
S.flexneri	ERR047296	MS0021	HCMC	1995	-	Susceptible
S.flexneri	ERR047297	MS0022	HCMC	1995	-	Resistant
S.flexneri	ERR047298	MS0025	HCMC	1995	-	Susceptible
S.flexneri	ERR047299	MS0026	HCMC	1995	-	Susceptible
S.flexneri	ERR047300	MS0029	HCMC	1995	-	Susceptible
S.flexneri	ERR047301	MS0036	HCMC	1995	-	Susceptible
S.flexneri	ERR047302	MS0038	HCMC	1995	-	Susceptible
S.flexneri	ERR047303	MS0041	HCMC	1995	-	Susceptible
S.flexneri	ERR047305	MS0047	HCMC	1995	-	Susceptible
S.flexneri	ERR047306	MS0050	HCMC	1995	-	Susceptible
S.flexneri	ERR047307	MS0052	HCMC	1995	-	Susceptible
S.flexneri	ERR047308	MS0055	HCMC	1995	-	Resistant
S.flexneri	ERR047309	MS0059	HCMC	1995	-	Susceptible
S.flexneri	ERR047310	10398	HCMC	2010	-	Resistant

S.flexneri	ERR047329	30078	HCMC	2009	-	Resistant
S.flexneri	ERR047333	30157	HCMC	2009	-	Resistant
S.flexneri	ERR047346	DE0010	HCMC	2000	-	Resistant
S.flexneri	ERR047347	DE0029	HCMC	2000	-	Susceptible
S.flexneri	ERR047348	DE0030	HCMC	2000	-	Resistant
S.flexneri	ERR047349	DE0042	HCMC	2000	-	Susceptible
S.flexneri	ERR047351	DE0090	HCMC	2000	-	Susceptible
S.flexneri	ERR047352	DE0119	HCMC	2000	-	Resistant
S.flexneri	ERR047353	DE0174	HCMC	2000	-	Resistant
S.flexneri	ERR047354	DE0175	HCMC	2000	-	Resistant
S.flexneri	ERR047355	DE0184	HCMC	2000	-	Resistant
S.flexneri	ERR047357	DE0195	HCMC	2000	-	Susceptible
S.flexneri	ERR047358	DE0213	HCMC	2000	-	Resistant
S.flexneri	ERR047359	DE0224	HCMC	2000	-	Resistant
S.flexneri	ERR047360	DE0319	HCMC	2000	-	Susceptible
S.flexneri	ERR047361	DE0350	HCMC	2000	-	Susceptible
S.flexneri	ERR047362	DE0366	HCMC	2000	-	Susceptible
S.flexneri	ERR047363	DE0438	HCMC	2000	-	Susceptible
S.flexneri	ERR047364	DE0457	HCMC	2000	-	Resistant
S.flexneri	ERR047365	DE0569	HCMC	2000	-	Resistant
S.flexneri	ERR047366	DE0578	HCMC	2001	-	Susceptible
S.flexneri	ERR047367	DE0614	HCMC	2001	-	Susceptible
S.flexneri	ERR047368	DE0670	HCMC	2001	-	Resistant
S.flexneri	ERR047369	DE0692	HCMC	2001	-	Resistant
S.flexneri	ERR047370	DE0712	HCMC	2001	-	Resistant
S.flexneri	ERR047371	DE0721	HCMC	2001	-	Resistant
S.flexneri	ERR047372	DE0804	HCMC	2001	-	Resistant
S.flexneri	ERR047373	DE0807	HCMC	2001	-	Resistant
S.flexneri	ERR047374	DE0874	HCMC	2002	-	Susceptible
S.flexneri	ERR047376	DE1173	HCMC	2002	-	Susceptible
S.flexneri	ERR047377	DE1174	HCMC	2002	-	Resistant
S.flexneri	ERR047378	DE1179	HCMC	2002	-	Resistant
S.flexneri	ERR047379	DE1194	HCMC	2002	-	Resistant
S.flexneri	ERR047380	DE1244	HCMC	2002	-	Resistant
S.flexneri	ERR047381	DE1252	HCMC	2002	-	Resistant
S.flexneri	ERR047382	DE1279	HCMC	2002	-	Resistant
S.flexneri	ERR047383	DE1342	HCMC	2002	-	Resistant
S.flexneri	ERR047384	DE1455	HCMC	2002	-	Susceptible
S.flexneri	ERR047385	DE1461	HCMC	2002	-	Susceptible
S.flexneri	ERR047386	DE1496	HCMC	2002	-	Resistant
S.flexneri	ERR047387	DE1512	HCMC	2002	-	Susceptible
S.flexneri	ERR047388	EG0002	HCMC	2006	-	Resistant

S.flexneri	ERR047389	EG0006	HCMC	2006	-	Susceptible
S.flexneri	ERR047390	EG0055	HCMC	2006	-	Susceptible
S.flexneri	ERR047391	EG0059	HCMC	2006	-	Susceptible
S.flexneri	ERR047392	EG0085	HCMC	2007	-	Susceptible
S.flexneri	ERR047393	EG0134	HCMC	2007	-	Resistant
S.flexneri	ERR047394	EG0198	HCMC	2008	-	Resistant
S.flexneri	ERR047395	EG0257	HCMC	2008	-	Resistant
S.flexneri	ERR047396	EG0302	HCMC	2006	-	Susceptible
S.flexneri	ERR047397	EG0305	HCMC	2006	-	Susceptible
S.flexneri	ERR047398	EG0319	HCMC	2006	-	Resistant
S.flexneri	ERR047399	EG0329	HCMC	2006	-	Resistant
S.flexneri	ERR047400	EG0380	HCMC	2007	-	Resistant
S.flexneri	ERR047401	EG0387	HCMC	2007	-	Susceptible
S.flexneri	ERR047402	EG0403	HCMC	2007	-	Resistant
S.flexneri	ERR047403	EG0412	HCMC	2007	-	Resistant
S.flexneri	ERR047404	EG0419	HCMC	2007	-	Resistant
S.flexneri	ERR047405	EG0435	HCMC	2008	-	Resistant
S.flexneri	ERR047406	EG0449	HCMC	2008	-	Resistant
S.flexneri	ERR047407	EG0461	HCMC	2008	-	Resistant
S.flexneri	ERR047408	EG0469	HCMC	2008	-	Resistant
S.flexneri	ERR047409	EG0471	HCMC	2008	-	Resistant
S.flexneri	ERR047410	EG0474	HCMC	2008	-	Resistant
S.flexneri	ERR047429	HUE03	Hue	2008	-	Resistant
S.flexneri	ERR047430	HUE04	Hue	2008	-	Resistant
S.flexneri	ERR047432	HUE06	Hue	2009	-	Resistant
S.flexneri	ERR047433	HUE07	Hue	2009	-	Resistant
S.flexneri	ERR047434	HUE08	Hue	2009	-	Resistant
S.flexneri	ERR047435	HUE09	Hue	2009	-	Resistant
S.flexneri	ERR047436	HUE10	Hue	2009	-	Resistant
S.flexneri	ERR047437	HUE12	Hue	2009	-	Resistant
S.flexneri	ERR047438	HUE13	Hue	2009	-	Resistant
S.flexneri	ERR047439	HUE14	Hue	2009	-	Resistant
S.flexneri	ERR047440	HUE15	Hue	2009	-	Resistant
S.flexneri	ERR047443	HUE18	Hue	2009	-	Resistant
S.flexneri	ERR048230	MS0061	HCMC	1995	-	Susceptible
S.flexneri	ERR048231	MS0064	HCMC	1995	-	Susceptible
S.flexneri	ERR048232	MS0076	HCMC	1995	-	Susceptible
S.flexneri	ERR048233	MS0078	HCMC	1995	-	Susceptible
S.flexneri	ERR048234	MS0084	HCMC	1996	-	Susceptible
S.flexneri	ERR048235	MS0085	HCMC	1996	-	Susceptible
S.flexneri	ERR048236	MS0087	HCMC	1996	-	Susceptible
S.flexneri	ERR048237	MS0091	HCMC	1996	-	Resistant

S.flexneri	ERR048238	MS0093	HCMC	1996	-	Susceptible
S.flexneri	ERR048240	MS0099	HCMC	1996	-	Susceptible
S.flexneri	ERR048241	MS0100	HCMC	1996	-	Susceptible
S.flexneri	ERR048242	MS0101	HCMC	1996	-	Resistant
S.flexneri	ERR048243	MS0105	HCMC	1996	-	Susceptible
S.flexneri	ERR048244	MS0114	HCMC	1996	-	Susceptible
S.flexneri	ERR048245	MS0118	HCMC	1996	-	Susceptible
S.flexneri	ERR048246	MS0125	HCMC	1996	-	Susceptible
S.flexneri	ERR048247	MS0126	HCMC	1996	-	Susceptible
S.flexneri	ERR048248	MS0130	HCMC	1996	-	Susceptible
S.flexneri	ERR048249	MS0131	HCMC	1996	-	Resistant
S.flexneri	ERR048250	MS0134	HCMC	1996	-	Susceptible
S.flexneri	ERR048251	MS0135	HCMC	1996	-	Susceptible
S.flexneri	ERR049133	10013	HCMC	2010	-	Resistant
S.flexneri	ERR049152	10262	HCMC	2010	-	Resistant
S.sonnei	ERR047223	HUE50	Hue	2009	Resistant	Resistant
S.sonnei	ERR047228	HUE57	Hue	2010	Susceptible	Susceptible
S.sonnei	ERR047229	HUE58	Hue	2010	Resistant	Resistant
S.sonnei	ERR047232	HUE62	Hue	2010	Susceptible	Susceptible
S.sonnei	ERR047234	HUE64	Hue	2010	Susceptible	Susceptible
S.sonnei	ERR047237	HUE67	Hue	2010	Susceptible	Susceptible
S.sonnei	ERR047238	HUE68	Hue	2010	Susceptible	Susceptible
S.sonnei	ERR047240	KH02	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047241	KH04	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047242	KH05	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047243	KH06	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047244	KH07	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047245	KH09	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047246	KH10	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047247	KH11	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047248	KH12	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047249	KH13	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047250	KH14	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047251	KH15	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047252	KH16	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047253	KH17	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047254	KH18	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047255	KH19	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047256	KH20	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047257	KH21	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047258	KH23	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047259	KH24	Khanh	2009	Resistant	Resistant

S.sonnei	ERR047260	KH25	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047261	KH26	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047262	KH27	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047263	KH28	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047264	KH29	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047265	KH30	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047266	KH32	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047267	KH33	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047268	KH34	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047269	KH35	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047270	KH37	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047271	KH38	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047273	KH40	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047274	KH41	Khanh	2009	Resistant	Resistant
S.sonnei	ERR047275	KH42	Khanh	2010	Susceptible	Susceptible
S.sonnei	ERR047276	KH43	Khanh	2010	Resistant	Resistant
S.sonnei	ERR047277	KH45	Khanh	2010	Resistant	Resistant
S.sonnei	ERR047283	KH53	Khanh	2010	Resistant	Resistant
S.sonnei	ERR047284	KH54	Khanh	2010	Resistant	Resistant
S.sonnei	ERR047285	KH55	Khanh	2010	Resistant	Resistant
S.sonnei	ERR047286	KH57	Khanh	2010	Resistant	Resistant
S.sonnei	ERR047304	MS0043	HCMC	1995	Susceptible	Susceptible
S.sonnei	ERR047311	20006	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047312	20021	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047313	20023	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047314	20037	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047315	20070	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047316	20094	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047317	20228	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047318	20261	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047319	20263	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047320	20343	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047321	30003	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047322	30008	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047323	30010	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047324	30037	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047325	30054	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047326	30059	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047327	30071	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047328	30073	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047330	30100	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047331	30112	HCMC	2009	Resistant	Resistant

S.sonnei	ERR047332	30124	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047334	30162	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047335	30164	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047336	30169	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047337	30172	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047338	30174	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047339	30233	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047340	30293	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047341	30366	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047342	30371	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047343	30387	HCMC	2010	Resistant	Resistant
S.sonnei	ERR047344	30450	HCMC	2010	Resistant	Resistant
S.sonnei	ERR047345	30451	HCMC	2010	Resistant	Resistant
S.sonnei	ERR047411	EG1014	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047412	EG1015	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047413	EG1016	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047414	EG1017	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047415	EG1018	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047416	EG1019	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047417	EG1020	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047418	EG1021	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047419	EG1022	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047420	EG1023	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047421	EG1024	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047422	EG1025	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047423	EG1026	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047424	EG1027	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047425	EG1028	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047426	EG1029	HCMC	2009	Resistant	Resistant
S.sonnei	ERR047427	HUE01	Hue	2008	Susceptible	Susceptible
S.sonnei	ERR047428	HUE02	Hue	2008	Susceptible	Susceptible
S.sonnei	ERR047431	HUE05	Hue	2008	Susceptible	Susceptible
S.sonnei	ERR047441	HUE16	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047442	HUE17	Hue	2009	Resistant	Resistant
S.sonnei	ERR047444	HUE19	Hue	2009	Resistant	Resistant
S.sonnei	ERR047445	HUE20	Hue	2009	Resistant	Resistant
S.sonnei	ERR047446	HUE21	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047447	HUE22	Hue	2009	Resistant	Resistant
S.sonnei	ERR047448	HUE23	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047449	HUE24	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047450	HUE25	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047451	HUE26	Hue	2009	Susceptible	Susceptible

S.sonnei	ERR047452	HUE27	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047453	HUE29	Hue	2009	Resistant	Resistant
S.sonnei	ERR047454	HUE30	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047455	HUE31	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047456	HUE32	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047457	HUE33	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047458	HUE34	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047461	HUE40	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047462	HUE42	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047463	HUE43	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047464	HUE46	Hue	2009	Resistant	Resistant
S.sonnei	ERR047465	HUE47	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR047466	HUE48	Hue	2009	Susceptible	Susceptible
S.sonnei	ERR048239	MS0094	HCMC	1996	Susceptible	Susceptible
S.sonnei	ERR049134	10014	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049135	10021	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049136	10031	HCMC	2009	Susceptible	Susceptible
S.sonnei	ERR049137	10035	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049138	10060	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049139	10063	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049140	10071	HCMC	2009	Susceptible	Susceptible
S.sonnei	ERR049141	10073	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049142	10083	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049143	10093	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049144	10102	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049145	10111	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049146	10115	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049147	10134	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049148	10135	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049149	10152	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049150	10159	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049151	10188	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049153	10263	HCMC	2009	Resistant	Resistant
S.sonnei	ERR049154	10320	HCMC	2010	Resistant	Resistant
S.sonnei	ERR049155	10365	HCMC	2010	Resistant	Resistant

Supplementary Table 4.1.Vietnamese *S. sonnei* and *S. flexenri* isolates used in chapters 4 and 5 of this thesis. The date of isolation, location, accession number and phenotypic resistance to nalidixic acid (a quinolone) in *S. sonnei* are shown

Supplementary Material – chapter 5

Species name	Run accession	Strain Name	Lineage	Serotype	Year of Isolation	Country of isolation
Shigella flexneri	ERR042796	Sh02-1115	3	Xv	2002	Gabon
Shigella flexneri	ERR042799	Sh02-3648	3	2a	2002	Mayotte
Shigella flexneri	ERR042801	Sh02-8484	3	2a	2002	Haiti
Shigella flexneri	ERR042802	Sh03-0850	3	2a	2003	Haiti
Shigella flexneri	ERR042804	Sh03-2590	3	2a	2003	Madagascar
Shigella flexneri	ERR042806	Sh04-2878	3	2b	2004	Morocco
Shigella flexneri	ERR042807	Sh04-4915	1	1b	2004	Algeria
Shigella flexneri	ERR042810	Sh04-9462	1	1b	2004	Cameroon
Shigella flexneri	ERR042812	Sh05-1382	1	4av	2005	Mali
Shigella flexneri	ERR042813	Sh05-1396	1	1b	2005	Chad
Shigella flexneri	ERR042814	Sh05-1631	3	2a	2005	Dominican Republic
Shigella flexneri	ERR042824	Sh05-9228	3	2a	2005	French Guiana
Shigella flexneri	ERR042827	Sh06-1976	7	4a	2006	Ivory Coast
Shigella flexneri	ERR042828	Sh06-2028	1	1b	2006	Guinea
Shigella flexneri	ERR042829	Sh06-3638	2	3a	2006	Niger
Shigella flexneri	ERR042830	Sh06-7888	3	2a	2006	Algeria
Shigella flexneri	ERR042832	Sh07-0130	6	Y	2007	Madagascar
Shigella flexneri	ERR042836	Sh07-4848	2	3a	2007	Togo
Shigella flexneri	ERR042838	Sh07-5519	7	Yv	2007	Cameroon
Shigella flexneri	ERR042840	Sh07-6497	3	Yv	2007	Ivory Coast
Shigella flexneri	ERR042842	Sh07-7746	3	2a	2007	Guadeloupe
Shigella flexneri	ERR042843	Sh08-0350	4	3a	2008	Dominican Republic
Shigella flexneri	ERR042845	Sh08-1372	3	Xv	2008	Senegal

Shigella flexneri	ERR042846	Sh08-2354	1	1b	2008	Burkina Faso
Shigella flexneri	ERR042848	Sh08-3483	2	3a	2008	French Guiana
Shigella flexneri	ERR042851	Sh09-0167	1	1b	2009	Senegal
Shigella flexneri	ERR042854	Sh09-2087	3	2a	2009	Mayotte
Shigella flexneri	ERR042855	Sh09-4409	2	3a	2009	Togo
Shigella flexneri	ERR042858	Sh09-5516	7	4av	2009	Congo
Shigella flexneri	ERR042860	Sh09-6787	3	2a	2009	Comores
Shigella flexneri	ERR042863	Sh09-9407	3	Yv	2009	Mali
Shigella flexneri	ERR047372	DE0804	7	4av	2001	Vietnam
Shigella flexneri	ERR047385	DE1461	2	3a	2002	Vietnam
Shigella flexneri	ERR047394	EG0198	2	3a	2008	Vietnam
Shigella flexneri	ERR048288	IB4235	6	Y	2000	India
Shigella flexneri	ERR048291	IB3350	1	1b	2002	Pakistan
Shigella flexneri	ERR048295	IB3391	3	2b	2002	Pakistan
Shigella flexneri	ERR048302	IB0711	3	2a	1999	Korea
Shigella flexneri	ERR048303	IB0712	3	2a	2003	Korea
Shigella flexneri	ERR048319	IB0034	3	1a	2002	China
Shigella flexneri	ERR048321	IB0037	3	2a	2002	China
Shigella flexneri	ERR048322	IB0039	3	2a	2002	Philippines
Shigella flexneri	ERR048323	IB0043	3	2a	2002	Philippines
Shigella flexneri	ERR048324	IB0045	3	2a	2002	Taiwan
Shigella flexneri	ERR048327	IB0068	3	2a	2003	Taiwan
Shigella flexneri	ERR048328	IB1716	7	Yv	2002	Vietnam
Shigella flexneri	ERR048339	IB0064	3	2b	2002	Sri Lanka
Shigella flexneri	ERR126958	370076	3	2a	2009	South Africa (KwaZulu-Natal)
Shigella flexneri	ERR126963	319743	2	3a	2009	South Africa (Eastern Cape)
Shigella flexneri	ERR126966	356458	2	3a	2009	South Africa (Free State)

Shigella flexneri	ERR126986	441003	3	2a	2010	South Africa (Western Cape)
Shigella flexneri	ERR126987	466310	3	2a	2010	South Africa (KwaZulu-Natal)
Shigella flexneri	ERR126992	428153	2	3a	2010	South Africa (Eastern Cape)
Shigella flexneri	ERR127012	579748	3	2a	2011	South Africa (Gauteng)
Shigella flexneri	ERR127014	527649	3	2a	2011	South Africa (Western Cape)
Shigella flexneri	ERR127015	535661	3	2a	2011	South Africa (Gauteng)
Shigella flexneri	ERR200365	K-9282	3	2a	2009	Bangladesh
Shigella flexneri	ERR200367	K-9324	3	2a	2009	Bangladesh
Shigella flexneri	ERR217024	K-9165	6	Y	2009	Bangladesh
Shigella flexneri	ERR217029	K-8263	6	Yv	2007	Bangladesh
Shigella sonnei	ERR024064	20061758	II	NA	2006	Dominican Republic
Shigella sonnei	ERR024068	20041367	III	NA	2004	UK
Shigella sonnei	ERR024071	20040880	III	NA	2004	Sri Lanka
Shigella sonnei	ERR024077	CS20	III	NA	2002	Brazil
Shigella sonnei	ERR024081	20021122	III	NA	2002	UK
Shigella sonnei	ERR024083	19984123	III	NA	1998	Mexico
Shigella sonnei	ERR024084	20040489	I	NA	2004	UK
Shigella sonnei	ERR024087	20011685	III	NA	2001	UK
Shigella sonnei	ERR024090	20062087	III	NA	2006	Egypt_Tunisia
Shigella sonnei	ERR024096	CS7	III	NA	2000	Brazil
Shigella sonnei	ERR024097	CS8	III	NA	2000	Brazil
Shigella sonnei	ERR024604	#04 1191	II	NA	2004	Tanzania
Shigella sonnei	ERR024605	#06 6470	II	NA	2006	Haiti
Shigella sonnei	ERR024606	#07 4369	I	NA	2007	France
Shigella sonnei	ERR024607	#05 5623	II	NA	2005	Morocco
Shigella sonnei	ERR024608	#06 0108	III	NA	2006	French Guiana
Shigella sonnei	ERR024609	#06 2542	III	NA	2006	Burkina Faso

Shigella sonnei	ERR024610	#06 5179	III	NA	2006	Senegal
Shigella sonnei	ERR024611	#06 5387	III	NA	2006	Morocco
Shigella sonnei	ERR024612	#06 5623	III	NA	2006	Morocco
Shigella sonnei	ERR024614	#06 6396	III	NA	2006	France
Shigella sonnei	ERR024616	IB1	III	NA	2003	Korea
Shigella sonnei	ERR024617	IB690	III	NA	2000	Korea
Shigella sonnei	ERR024618	IB691	III	NA	1999	Korea
Shigella sonnei	ERR024620	IB2	III	NA	2003	Korea
Shigella sonnei	ERR024621	IB3	III	NA	2003	Korea
Shigella sonnei	ERR024622	IB10	III	NA	2003	Korea
Shigella sonnei	ERR025704	IB2000	III	NA	2003	Vietnam
Shigella sonnei	ERR025706	IB2008	III	NA	2003	Vietnam
Shigella sonnei	ERR025708	IB2015	III	NA	2002	Vietnam
Shigella sonnei	ERR025709	IB3488	III	NA	2003	Pakistan
Shigella sonnei	ERR025710	IB3507	III	NA	2003	Pakistan
Shigella sonnei	ERR025711	IB3580	III	NA	2003	Pakistan
Shigella sonnei	ERR025713	IB2024	III	NA	2002	Vietnam
Shigella sonnei	ERR025714	IB2026	III	NA	2003	Vietnam
Shigella sonnei	ERR025717	IB3277	III	NA	2002	Pakistan
Shigella sonnei	ERR025718	IB3300	III	NA	2002	Pakistan
Shigella sonnei	ERR025719	IB3374	III	NA	2002	Pakistan
Shigella sonnei	ERR025748	#97 0044	III	NA	1997	New Caledonia
Shigella sonnei	ERR025758	#98 8743	III	NA	1998	French Guiana
Shigella sonnei	ERR025759	#03 6224	III	NA	2003	Senegal
Shigella sonnei	ERR025761	#98 9560	III	NA	1998	Madagascar
Shigella sonnei	ERR025762	#9810267	II	NA	1998	Madagascar
Shigella sonnei	ERR025763	#99 8911	II	NA	1999	France

Shigella sonnei	ERR025764	#00 2225	III	NA	2000	French Guiana
Shigella sonnei	ERR025765	#00 5827	I	NA	2000	Madagascar
Shigella sonnei	ERR025767	#03 1382	III	NA	2003	Israel
Shigella sonnei	ERR025768	#03 2222	III	NA	2003	Cuba
Shigella sonnei	ERR028671	20051272	III	NA	2005	Egypt
Shigella sonnei	ERR028677	20060018	III	NA	2006	Egypt
Shigella sonnei	ERR028681	20051541	III	NA	2005	Uzbekistan
Shigella sonnei	ERR028684	20061309	III	NA	2006	Egypt
Shigella sonnei	ERR028686	CS6	III	NA	2000	Brazil
Shigella sonnei	ERR028687	CS14	III	NA	2001	Brazil
Shigella sonnei	ERR028690	20031275	III	NA	2003	Iran
Shigella sonnei	ERR028693	20052631	III	NA	2005	Peru
Shigella sonnei	ERR028697	20040924	III	NA	2004	Kenya
Shigella sonnei	ERR028700	20071599	II	NA	2007	UK
Shigella sonnei	ERR028702	20010007	III	NA	2001	UK
Shigella sonnei	ERR028703	20062313	III	NA	2006	Nepal

Supplementary Table 5.1. Global *S. sonnei* and *S. flexenri* isolates used in chapter 5 of this thesis. Strain information including date of isolation, country of isolation, lineage and run accession number are shown.

.

HGC gene name	Function	Number of *S. sonnei* isolates	Number of *S. flexneri* isolates	HGC with the same gene name annotation (*S. sonnei* / *S. flexneri numbers*)
aaeB_2	p-hydroxybenzoic acid efflux pump subunit AaeB	146	4	aaeB_1 (139/135)
aarA	Rhomboid protease AarA	146	13	-
abgA	p-aminobenzoyl-glutamate hydrolase subunit A	144	10	abgA_1 (1/10)
abgB	p-aminobenzoyl-glutamate hydrolase subunit B	146	10	-
abgT_1	p-aminobenzoyl-glutamate transport protein	146	10	abgT_2 (146/0)
abgT_2	p-aminobenzoyl-glutamate transport protein	146	0	abgT_1 (146/10)
acoR	Acetoin dehydrogenase operon transcriptional activator AcoR	0	129	-
acrF_2	Multidrug export protein AcrF	146	19	acrF_1 (6/19), acrF_3 (146/13)
acrF_3	Multidrug export protein AcrF	146	13	acrF_1 (6/19), acrF_2 (146/19)
adhB	Alcohol dehydrogenase 2	0	134	-
adhE_3	Aldehyde-alcohol dehydrogenase	146	0	adhE_1 (145/52), adhE_2 (146/21), adhE_4 (139/15)
adhE_4	Aldehyde-alcohol dehydrogenase	139	15	adhE_1 (145/52), adhE_2 (146/21), adhE_3 (146/0)
adhT	Alcohol dehydrogenase	0	133	-
aldA_2	Lactaldehyde dehydrogenase	145	6	aldA (0/1), aldA_1 (145/126)
aldB	Aldehyde dehydrogenase B	0	133	aldB_1 (0/2), aldB_2 (0/2)
allA	Ureidoglycolate lyase	0	135	-
allB	Allantoinase	146	14	-
allR	HTH-type transcriptional repressor AllR	0	134	-
allS_5	HTH-type transcriptional activator AllS	139	14	allS_1 (94/0), allS_2 (144/122), allS_3 (145/133), allS_4 (145/134), allS_6 (145/58)
aplIR	Type-2 restriction enzyme AplI	146	0	-
araC_2	putative histidine kinase sensor domain protein	146	13	araC (146/135)
arcC1_2	Carbamate kinase 1	146	13	araC1_1 (146/130), araC1_3 (0/1)
arcC2	Carbamate kinase 2	146	0	-
argE_1	Acetylornithine deacetylase	145	13	argE (2/114), argE_2 (145/20)

argF	Ornithine carbamoyltransferase	145	13	-
aroG_2	Phospho-2-dehydro-3-deoxyheptonate aldolase, Phe-sensitive	146	3	aroG_1 (146/134)
aroP_2	Aromatic amino acid transport protein AroP	145	13	aroP_1 (146/134)
astC_1	Succinylornithine transaminase	19	130	astC (108/83), astC_2 (22/25)
atoA	Acetate CoA-transferase subunit beta	145	0	-
atoD	Acetate CoA-transferase subunit alpha	145	0	-
atoE	Short-chain fatty acids transporter	146	0	-
atzC	N-isopropylammelide isopropyl amidohydrolase	146	0	-
avtA	Valine--pyruvate aminotransferase	0	134	-
bcr_1	Bicyclomycin resistance protein	146	1	bcr (145/133)
bfd	Bacterioferritin-associated ferredoxin	0	135	-
bfr	Bacterioferritin	0	135	-
bglA_1	6-phospho-beta-glucosidase BglA	146	12	bgl_2 (146/135)
cadA_1	Lysine decarboxylase, inducible	146	0	cadA_2 (126/0), cadA_3 (144/0)
cadA_3	Lysine decarboxylase, inducible	144	0	cadA_1 (146/0), cadA_2 (
cadC_1	Transcriptional activator CadC	146	0	cadC_2 (145/0)
cadC_2	Transcriptional activator CadC	145	0	cadC_1 (146/0)
caiT_2	L-carnitine/gamma-butyrobetaine antiporter	146	13	caiT (146/122)
casA	CRISPR system Cascade subunit CasA	146	13	-
casC	CRISPR system Cascade subunit CasC	146	13	-
casD	CRISPR system Cascade subunit CasD	138	11	-
casE	CRISPR associated protein	144	13	-
ccmL	Carbon dioxide concentrating mechanism protein CcmL	146	0	-
cdaR_2	Carbohydrate diacid regulator	0	132	cdaR (146/119), cdaR_1 (0/5)
cdhB	Caffeine dehydrogenase subunit beta	143	0	-
cdsA_2	Phosphatidate cytidylyltransferase	146	1	cdsA_1 (146/135
chbC	N,N'-diacetylchitobiose permease IIC component	138	1	chbC_1 (4/106), chbC_2 (4/106)
citC_2	[Citrate [pro-3S]-lyase] ligase	146	13	citC_1 (145/122)

clpP1	ATP-dependent Clp protease proteolytic subunit 1	145	0	-
cytR_1	HTH-type transcriptional repressor CytR	145	0	cytR (146/135)
dacD	D-alanyl-D-alanine carboxypeptidase DacD precursor	146	0	-
ddpA	putative D,D-dipeptide-binding periplasmic protein DdpA precursor	143	13	-
ddpB	putative D,D-dipeptide transport system permease protein DdpB	146	13	-
ddrA	Diol dehydratase-reactivating factor alpha subunit	146	19	-
dgt_1	Deoxyguanosinetriphosphate triphosphohydrolase	146	0	dgt_2 (145/135)
dhaK	PTS-dependent dihydroxyacetone kinase, dihydroxyacetone-binding subunit DhaK	0	131	-
dhaL	PTS-dependent dihydroxyacetone kinase, ADP-binding subunit DhaL	0	130	-
dinB_2	DNA polymerase IV	145	5	dinB_1 (146/134), dinB_3 (140/91)
dinJ	Antitoxin DinJ	146	0	-
dlgD	2,3-diketo-L-gulonate reductase	0	133	-
dmlA	D-malate dehydrogenase [decarboxylating]	146	13	-
dmlR_6	HTH-type transcriptional regulator DmlR	138	4	dmlR_1 (146/135), dmlR_2 (139/120), dmlR_3 (145/116), dmlR_4 (146/133), dmlR_5 (123/87), dmlR_7 (140/42)
dnaC_2	DNA replication protein DnaC	146	10	dnaC_1 (146/135), dnaC_3 (146/135), dnaC_4 (0/4), dnaC_5 (1/4)
dnaJ_1	Chaperone protein DnaJ	13	133	dnaJ (142/105), dnaJ_2 (34/19), dnaJ_3 (1/0)
ecfT	Energy-coupling factor transporter transmembrane protein EcfT	0	134	-
ecpD_1	Chaperone protein EcpD precursor	146	13	ecpD_2 (146/105)
elfC_1	putative outer membrane usher protein ElfC precursor	146	18	elfC_2 (146/130), elfC_3 (4/88), elfC_4 (1/2), elfC_5 (140/15), elfC_6 (2/0)
elfC_5	putative outer membrane usher protein ElfC precursor	140	15	elfC_1 (146/18), elfC_2 (146/130), elfC_3 (4/88), elfC_4 (1/2), elfC_6 (2/0)
elfG_1	putative fimbrial-like protein ElfG precursor	145	19	elfG_2 (146/122), elfG_3 (142/109), elfG_4 (146/130)
envR	putative acrEF/envCD operon repressor	146	13	-
epsJ_2	putative glycosyltransferase EpsJ	146	13	epsJ (0/24), epsj_1 (146/31)

etfA_3	Electron transfer flavoprotein subunit alpha	0	132	eftA_1 (146/122), eftA_2 (146/114)
eutB_1	Ethanolamine ammonia-lyase heavy chain	146	17	eutB_2 (132/36)
exsA_2	Exoenzyme S synthesis regulatory protein ExsA	145	13	exsA (15/127), exsA_1 (131/29)
exuT_1	Hexuronate transporter	146	17	exuT_2 (145/133)
fecA_1	Fe(3) dicitrate transport protein FecA precursor	144	14	fecA (83/82), fecA_2 (69/4)
fhlA_3	Formate hydrogenlyase transcriptional activator	144	0	fhlA_1 (146/135), fhlA_2 (144/34), fhlA_4 (141/129)
fimA_1	Type-1 fimbrial protein, A chain precursor	146	19	fimA_2 (146/135), fimA_3 (0/113), fim_4 (0/13)
flgG_2	Flagellar basal-body rod protein FlgG	146	0	flgG_1 (146/135)
fliD	Flagellar hook-associated protein 2	0	131	fliD_1 (145/0), fliD_2 (146/0), fliD_3 (146/0)
fliD_1	Flagellar hook-associated protein 2	145	0	fliD (0/131), fliD_2 (146/0), fliD_3 (146/0)
fliD_2	Flagellar hook-associated protein 2	146	0	fliD (0/131), fliD_1 (145/0), fliD_3 (146/0)
fliD_3	Flagellar hook-associated protein 2	146	0	fliD (0/131), fliD_1 (145/0), fliD_2 (146/0)
fliZ_1	Regulator of sigma S factor FliZ	143	7	fliZ (0/125), fliZ_2 (133/3)
flr	Flavoredoxin	145	0	-
frmA	S-(hydroxymethyl)glutathione dehydrogenase	144	0	-
frmB	S-formylglutathione hydrolase FrmB	144	0	-
frmR	Transcriptional repressor FrmR	144	0	-
fucA_1	L-fuculose phosphate aldolase	144	14	fucA (1/2), fucA_2 (146/134)
gadW_1	HTH-type transcriptional regulator GadW	144	0	gadW_2 (145/135)
galF	UTP--glucose-1-phosphate uridylyltransferase	0	135	-
garR_3	2-hydroxy-3-oxopropionate reductase	146	13	garR_1 (143/86), garR_2 (146/134)
gbpR	HTH-type transcriptional regulator GbpR	146	10	-
glmU_1	Bifunctional protein GlmU	144	0	glmU_2 (146/135)
glpR_3	Glycerol-3-phosphate regulon repressor	146	13	glpR_1 (146/135), glpR_2 (146/135)
gltD_2	Glutamate synthase [NADPH] small chain	146	16	gltD_1 (145/134), gltD_3 (146/3), gltD_4 (126/91)
gltD_3	Glutamate synthase [NADPH] small chain	146	3	gltD_1 (145/134), gltD_2 (146/16), gltD_4 (126/91)
gltR_1	HTH-type transcriptional regulator GltR	146	12	gltR_2 (146/132)
gno	Gluconate 5-dehydrogenase	146	0	-
gntK_2	Thermoresistant gluconokinase	146	0	gntK (146/135)

gntR	HTH-type transcriptional regulator GntR	1	134	gntR_1 (145/2), gntR_2 (146/0)
gntR_1	HTH-type transcriptional regulator GntR	145	2	gntR (1/134), gntR_2 (146/0)
gntR_2	HTH-type transcriptional regulator GntR	146	0	gntR (1/134), gntR_1 (146/2)
group_10024	Fimbrial protein	146	3	-
group_10058	Transposase IS116/IS110/IS902 family protein	145	13	-
group_10059	alpha-mannosidase	146	13	-
group_10061	Transposase DDE domain protein	143	0	-
group_10083	Phage minor tail protein U	143	10	-
group_10084	Bacteriophage lambda head decoration protein D	144	10	-
group_10091	putative TonB-dependent receptor precursor	145	12	-
group_10431	Phage late control gene D protein (GPD)	145	0	-
group_10795	Fimbrial assembly protein (PilN)	0	132	-
group_11103	Fels-1 Prophage Protein-like protein	146	13	-
group_11105	IS2 transposase TnpB	141	0	-
group_11114	ImpA domain protein	146	0	-
group_11144	Prophage CP4-57 regulatory protein (AlpA)	145	0	-
group_11160	NIF3 (NGG1p interacting factor 3)	146	12	-
group_11161	Phage protein GP46	146	0	-
group_11162	Excisionase-like protein	140	1	-
group_11191	Phage minor tail protein	143	9	-

group_11194	Phage major capsid protein E	145	10	-
group_11195	Phage DNA packaging protein Nu1	142	10	-
group_11212	Baseplate J-like protein	146	13	-
group_11217	putative racemase	144	0	-
group_11240	IS1 transposase	146	11	-
group_12071	Type-F conjugative transfer system protein (TrbI_Ftype)	145	4	-
group_12373	Stress-induced bacterial acidophilic repeat motif protein	141	18	-
group_12512	putative fimbrial-like adhesin protein	146	0	-
group_12514	putative fimbrial-like adhesin protein	146	0	-
group_12516	putative fimbrial-like adhesin protein	146	0	-
group_12519	Transposase	139	0	-
group_12527	putative metallophosphoesterase	146	13	-
group_12539	Gene 25-like lysozyme	146	0	-
group_12541	Type VI secretion lipoprotein	145	0	-
group_12555	Ankyrin repeat protein	146	0	-
group_12562	N-glycosyltransferase	144	0	-
group_12589	transcriptional repressor DicA	146	0	-
group_12596	Limonene hydroxylase	146	13	-
group_12618	Mu-like prophage major head subunit gpT	144	0	-

group_12621	secY/secA suppressor protein	146	13	-
group_12622	lipoprotein	146	13	-
group_12633	Phage antitermination protein Q	146	0	-
group_12634	ORF6N domain protein	146	0	-
group_12649	SMP-30/Gluconolaconase/LRE-like region	144	13	-
group_12653	Glycosyltransferase family 9 (heptosyltransferase)	146	0	-
group_12657	Phage regulatory protein Rha (Phage_pRha)	146	10	-
group_12659	Caudovirales tail fibre assembly protein	146	0	-
group_12679	Invasin	146	0	-
group_12684	electron transport complex protein RnfC	146	0	-
group_12698	Phage antitermination protein Q	143	0	-
group_12702	prophage protein NinE	144	0	-
group_12711	Minor tail protein T	143	10	-
group_12712	Bacteriophage lambda minor tail protein (GpG)	143	0	-
group_12713	gpW	144	10	-
group_12716	ORF11CD3 domain protein	145	0	-
group_12737	Colicin E1 (microcin) immunity protein	146	0	-
group_12739	Gram-negative bacterial tonB protein	146	12	-
group_12753	CRISPR-associated protein Cse1 (CRISPR_cse1)	146	0	-

group_1896	reactivating factor for ethanolamine ammonia lyase	146	18	-
group_2722	AAA-like domain protein	138	4	-
group_2852	Serine dehydratase alpha chain	142	16	-
group_3820	Phage terminase large subunit	144	6	-
group_4311	Electron transfer flavoprotein-ubiquinone oxidoreductase	0	131	-
group_467	Phage portal protein	139	9	-
group_4838	Intracellular multiplication and human macrophage-killing	145	0	-
group_5229	putative E3 ubiquitin-protein ligase ipaH7.8	145	4	-
group_5348	putative transposase	2	132	-
group_5598	putative hydrolase	146	4	-
group_5599	putative hydrolase	0	133	-
group_5912	N-glycosyltransferase	142	0	-
group_5930	Poxvirus D5 protein-like protein	141	0	-
group_6277	WGR domain protein	0	129	-
group_6317	colanic acid biosynthesis protein	0	128	-
group_6775	Phage portal protein, lambda family	144	10	-
group_703	Lipoprotein Rz1 precursor	143	19	-
group_7337	Caudovirales tail fibre assembly protein	145	13	-
group_7455	PrkA AAA domain protein	144	13	-
group_7476	Phage tail sheath protein	146	16	-

group_7747	SpoVR family protein	144	16	-
group_8291	Phage integrase family protein	144	10	-
group_8341	Bacterial Ig-like domain (group 2)	144	8	-
group_8350	Transposase IS66 family protein	146	0	-
group_8603	TraE protein	146	4	-
group_8634	RNA 2'-phosphotransferase	0	133	-
group_8722	putative autotransporter precursor	0	134	-
group_8725	putative DNA-binding transcriptional regulator	142	9	-
group_9213	Bacteriophage lysis protein	140	18	-
group_9693	PGL/p-HBAD biosynthesis glycosyltransferase/MT3031	146	19	-
group_9953	Ethanolamine utilisation protein, EutH	146	19	-
gspA_3	General stress protein A	146	13	gspA_1 (146/122), gspA_2 (146/122), gspA_4 (146/13), gspA_5 (146/13)
gspA_4	General stress protein A	146	13	gspA_1 (146/122), gspA_2 (146/122), gspA_3 (146/13), gspA_5 (146/13)
gspA_5	General stress protein A	146	13	gspA_1 (146/122), gspA_2 (146/122), gspA_3 (146/13), gspA_4 (146/13)
gspB	Putative general secretion pathway protein B	146	0	-
gstA_1	Glutathione S-transferase GstA	146	17	gstA_2 (146/118)
gstB_1	Glutathione S-transferase GST-6.0	145	13	gstB (146/135)
guaD	Guanine deaminase	145	19	-
hcaR_3	Hca operon transcriptional activator	141	1	hcaR_1 (145/129), hcaR_2 (146/134)
hcpA_1	Major exported protein	146	0	hcpA (0/7), hcpA_2 (146/0)
hcpA_2	Major exported protein	146	0	hcpA (0/7), hcpA_1 (146/0)
hicB	Antitoxin HicB	146	8	-

hilA	Transcriptional regulator HilA	146	0	-
hipA	Serine/threonine-protein kinase HipA	146	0	-
hipB	Antitoxin HipB	146	0	-
hscC_1	Chaperone protein HscC	2	129	hscC (144/80), hscC_2 (0/4)
htrE_1	Outer membrane usher protein HtrE precursor	146	14	htrE (146/119), htrE_2 (115/0), htrE_3 (146/0), htrE_4 (146/13), htrE_5 (146/0), htrE_6 (146/5)
htrE_3	Outer membrane usher protein HtrE precursor	146	0	htrE (146/119), htrE_1 (146/14), htrE_2 (115/0), htrE_4 (146/13), htrE_5 (146/0), htrE_6 (146/5)
htrE_4	Outer membrane usher protein HtrE precursor	146	13	htrE (146/119), htrE_1 (146/14), htrE_2 (115/0), htrE_3 (146/0), htrE_5 (146/0), htrE_6 (146/5)
htrE_5	Outer membrane usher protein HtrE precursor	146	0	htrE (146/119), htrE_1 (146/14), htrE_2 (115/0), htrE_3 (146/0), htrE_4 (146/13), htrE_6 (146/5)
htrE_6	Outer membrane usher protein HtrE precursor	146	5	htrE (146/119), htrE_1 (146/14), htrE_2 (115/0), htrE_3 (146/0), htrE_4 (146/13), htrE_5 (146/0)
hyuA_1	D-phenylhydantoinase	146	1	hyuA_2 (146/13)
hyuA_2	D-phenylhydantoinase	146	13	hyuA_1 (146/1)
iadA	Isoaspartyl dipeptidase	0	132	iadA_1 (0/2), iadA_2 (0/1)
idnD	L-idonate 5-dehydrogenase (NAD(P)( ))	146	1	-
idnT	Gnt-II system L-idonate transporter	146	0	-
intA_4	hypothetical protein	0	135	intA (82/1), intA_1 (54/113), intA_2 (83/135), intA_3 (1/7)
iraD	Anti-adapter protein IraD	0	131	-
lacY_2	Lactose permease	146	0	lacY_1 (125/0)
leuE	Leucine efflux protein	146	13	leuE_1 (0/11), leuE_2 (0/1)
loiP_2	Metalloprotease LoiP precursor	146	15	loiP_1 (146/135)
lptF	Lipopolysaccharide export system permease protein LptF	145	18	lptF_1 (0/117), lptF_2 (0/120)
lysU	Lysine--tRNA ligase, heat inducible	146	0	-
malS	Alpha-amylase precursor	0	134	malS_1 (0/3), malS (0/1)
matA	HTH-type transcriptional regulator MatA	146	10	-
mazE	Antitoxin MazE	146	13	mazE_1 (0/1)
mazF	mRNA interferase MazF	146	13	mazF_1 (0/1)
mcbR_1	HTH-type transcriptional regulator McbR	144	0	mcbR (69/101), mcbR_2 (56/0)

mdh_1	NAD-dependent methanol dehydrogenase	146	19	mdh (146/135)
merA	Mercuric reductase	146	4	-
mhpC	2-hydroxy-6-oxononadienedioate/2-hydroxy-6-oxononatrienedioate hydrolase	144	0	mhpC_1 (72/0)
mhpD	2-keto-4-pentenoate hydratase	143	0	-
mhpF	Acetaldehyde dehydrogenase	144	0	-
mkaC	Virulence genes transcriptional activator	146	18	-
mngB_3	Mannosylglycerate hydrolase	146	0	mngB_1 (143/40), mngB_2 (145/116)
mocA	Molybdenum cofactor cytidylyltransferase	145	13	-
mshB	1D-myo-inositol 2-acetamido-2-deoxy-alpha-D-glucopyranoside deacetylase	144	0	-
mtaD	5-methylthioadenosine/S-adenosylhomocysteine deaminase	145	0	-
mtlR_1	Mannitol operon repressor	0	135	mtlR (89/72)
murP_1	PTS system N-acetylmuramic acid-specific EIIBC component	139	10	murP 4/105), murP_2 (145/18)
murP_2	PTS system N-acetylmuramic acid-specific EIIBC component	145	18	murP 4/105), murP_1 (139/10)
nadA	Quinolinate synthase A	0	135	nadA_1 (146/1), nadA_2 (140/0)
nadA_1	Quinolinate synthase A	146	1	nadA (0/135), nadA_2 (146/0)
nadA_2	Quinolinate synthase A	146	0	nadA (0/135), nadA_1 (146/1)
nanS_1	hypothetical protein	139	18	nanS (146/132), nanS_2 (74/31), nanS_3 (0/104)
narV	Respiratory nitrate reductase 2 gamma chain	144	0	-
narW	putative nitrate reductase molybdenum cofactor assembly chaperone NarW	145	0	-
nhoA	N-hydroxyarylamine O-acetyltransferase	143	0	-
ompN_3	Outer membrane protein N precursor	146	0	ompN_1 (1/114), ompN_2 (146/56)
outO	Type 4 prepilin-like proteins leader peptide-processing enzyme	19	133	-
pac	Penicillin G acylase precursor	145	13	pac_1 (10/8)
papC_4	Outer membrane usher protein PapC precursor	145	6	papC (0/11), papC_1 (144/99), papC_2 (145/49), papC_3 (135/121), papC_5 (144/1), papC_6 (146/0)
papC_5	Outer membrane usher protein PapC precursor	144	1	papC (0/11), papC_1 (144/99), papC_2 (145/49), papC_3 (135/121), papC_4 (145/13), papC_6 (146/0)

papC_6	Outer membrane usher protein PapC precursor	146	0	papC (0/11), papC_1 (144/99), papC_2 (145/49), papC_3 (135/121), papC_4 (145/13), papC_5 (144/1)
pbpX	Putative penicillin-binding protein PbpX	146	13	-
pcaK	4-hydroxybenzoate transporter PcaK	144	0	-
pduA_1	Propanediol utilization protein PduA	146	0	pduA_2 (145/0)
pduA_2	Propanediol utilization protein PduA	145	0	pduA_1 (146/0)
pduB	Propanediol utilization protein PduB	146	0	-
pduC	Propanediol dehydratase large subunit	146	0	-
pduD	Propanediol dehydratase medium subunit	145	0	-
pduE	Propanediol dehydratase small subunit	146	0	-
pduF	Propanediol diffusion facilitator	146	0	-
pduL	Phosphate propanoyltransferase	146	0	-
pduU	Propanediol utilization protein PduU	146	0	-
pduV_1	Propanediol utilization protein PduV	146	1	pduV_2 (136/135)
pemK	mRNA interferase PemK	0	135	pemK (1/3)
pimB	GDP-mannose-dependent alpha-(1-6)-phosphatidylinositol monomannoside mannosyltransferase	0	133	-
pinR_1	Putative DNA-invertase from lambdoid prophage Rac	146	1	pinR (0/28), pinR_2 (146/0)
pinR_2	Putative DNA-invertase from lambdoid prophage Rac	146	0	pinR (0/28), pinR_1 (146/1)
potA_3	Spermidine/putrescine import ATP-binding protein PotA	146	0	potA_1 (146/135), potA_2 (146/135)
potF_2	Putrescine-binding periplasmic protein precursor	145	0	potF (2/71), potF_1 (145/64)
pptA	Tautomerase PptA	145	0	-
preT_3	NAD-dependent dihydropyrimidine dehydrogenase subunit PreT	143	0	preT_1 (145/26), preT_2 (146/115)
pstS1	Phosphate-binding protein PstS 1 precursor	146	12	-
pucA	putative xanthine dehydrogenase subunit A	146	13	-
putA	Bifunctional protein PutA	145	12	putA_2 (1/1)
putP	Sodium/proline symporter	145	13	-

puuE_2	4-aminobutyrate aminotransferase PuuE	141	11	puuE (2/37), puuE_1 (144/89)
racX	putative amino-acid racemase	0	134	-
recQ_1	ATP-dependent DNA helicase RecQ	146	1	recQ (146/135), recQ_2 (0/1)
relE2	Toxin RelE2	146	1	relE2_1 (18/3)
rfaL	O-antigen ligase	146	13	-
rfbB	dTDP-glucose 4,6-dehydratase	0	132	rfbB_1 (143/118)
rfbC	dTDP-4-dehydrorhamnose 3,5-epimerase	0	134	-
rfbD	dTDP-4-dehydrorhamnose reductase	0	133	-
rhaR_2	HTH-type transcriptional activator RhaR	146	16	rhaR_1 (146/135), rhaR_3 (141/119)
rhaT_1	L-rhamnose-proton symporter	2	135	rhaT (144/39), rhaT_2 (0/5)
rhmD	L-rhamnonate dehydratase	146	13	rhmD_1 (1/0)
rhmR	putative HTH-type transcriptional regulator RhmR	145	13	-
rhsA_1	putative deoxyribonuclease RhsA	145	0	rhsA_2 (143/0), rhsA_3 (6/0), rhsA_4 (5/1)
rhsA_2	putative deoxyribonuclease RhsA	143	0	rhsA_1 (145/0), rhsA_3 (6/0), rhsA_4 (5/1)
rhsC_2	hypothetical protein	142	12	rhsC_1 (117/60), rhsC_3 (78/5), rhsC_4 (136/4), rhsC_5 (75/3), rhsC_6 (138/10)
rhsC_6	hypothetical protein	138	10	rhsC_1 (117/60), rhsC_2 (142/12), rhsC_3 (78/5), rhsC_4 (136/4), rhsC_5 (75/3)
rmlA1	Glucose-1-phosphate thymidylyltransferase 1	0	134	-
rpiB	Ribose-5-phosphate isomerase B	146	13	-
rutA_2	Pyrimidine monooxygenase RutA	141	1	rutA (116/118), rutA_1 (27/1)
rutC_1	Putative aminoacrylate peracid reductase RutC	142	0	rutC (3/111), rutC_2 (144/7)
rutC_2	Putative aminoacrylate peracid reductase RutC	144	7	rutC (3/111), rutC_1 (142/0)
salL	Adenosyl-chloride synthase	146	13	-
sbmC	DNA gyrase inhibitor	146	0	-
sfaG	S-fimbrial protein subunit SfaG precursor	0	131	-
sfaH_1	S-fimbrial protein subunit SfaH	0	133	sfaH (146/122), sfaH_2 (0/13)
sfaS	S-fimbrial adhesin protein SfaS precursor	0	134	-
sicA	Chaperone protein SicA	146	0	-
sinR	HTH-type transcriptional regulator SinR	145	8	-

sorC_1	Sorbitol operon regulator	146	10	sorC (0/23), sorC_2 (145/95)
sppA_2	Putative signal peptide peptidase SppA	141	10	sppA (0/106), sppA_1 (146/54)
srlD_2	Sorbitol-6-phosphate 2-dehydrogenase	143	14	srlD_1 (143/120)
srlR_4	Glucitol operon repressor	138	14	srlR_1 (146/135), srlR_2 (146/135), srlR_3 (146/135)
ssb_1	Single-stranded DNA-binding protein	146	19	ssb_2 (111/122), ssb_4 (0/1)
sucD_1	Succinyl-CoA ligase [ADP-forming] subunit alpha	146	0	sucD (146/135), sucD_2 (1/0)
tabA_3	Toxin-antitoxin biofilm protein TabA	144	0	tabA_1 (146/135), tabA_2 (145/135)
taqIM	Modification methylase TaqI	146	0	-
tehB_1	putative S-adenosyl-L-methionine-dependent methyltransferase TehB	146	13	tehB_2 (146/128)
thlA_1	Acetyl-CoA acetyltransferase	143	0	thlA (0/39), thlA_2 (142/111)
tibA_2	Adhesin/invasin TibA autotransporter precursor	145	13	tibA (0/3), tibA_1 (146/43)
topB	DNA topoisomerase 3	141	13	topB_1 (20/55), topB_2 (6/122)
torA_2	Trimethylamine-N-oxide reductase 1 precursor	145	11	torA_1 (144/112)
torI	Response regulator inhibitor for tor operon	146	0	-
treA_2	Periplasmic trehalase precursor	0	128	treA (138/130), treA_1 (0/1)
virF	Virulence regulon transcriptional activator VirF	146	0	virF_1 (65/129), virF_2 (0/122)
wcaJ	UDP-glucose:undecaprenyl-phosphate glucose-1-phosphate transferase	0	132	-
wzxC_1	Lipopolysaccharide biosynthesis protein WzxC	0	134	wzxC_2 (0/134)
wzxC_2	Lipopolysaccharide biosynthesis protein WzxC	0	134	wzxC_1 (0/134)
xdhA_2	Xanthine dehydrogenase molybdenum-binding subunit	146	1	xdhA (1/26), xdhA_1 (145/94)
xylF	D-xylose-binding periplasmic protein precursor	0	135	-
xylG	2-hydroxymuconic semialdehyde dehydrogenase	1	134	-
xylH	Xylose transport system permease protein XylH	0	132	-
xylR_1	Xylose operon regulatory protein	0	134	xylR_2 (0/108)
yafQ	mRNA interferase YafQ	146	0	-
yagU	Inner membrane protein YagU	146	13	-
yahK	Aldehyde reductase YahK	145	0	-

ybbY	Putative purine permease YbbY	146	14	-
ybdO_1	putative HTH-type transcriptional regulator YbdO	0	135	ybdO (146/135)
ybhA	Pyridoxal phosphate phosphatase YbhA	0	135	-
ybhN_2	Inner membrane protein YbhN	143	19	ybhN (3/79), ybhN_1 (145/41)
ycgF_1	Blue light- and temperature-regulated antirepressor YcgF	145	4	ycgF (0/13), ycgF_2 (146/21)
ycjP_1	Inner membrane ABC transporter permease protein YcjP	144	2	ycjP (0/1), ycjP_2 (145/117)
ydcR	putative HTH-type transcriptional regulator YdcR	146	0	-
ydcU_1	Inner membrane ABC transporter permease protein YdcU	144	0	ydcU_2 (144/101)
yddE	putative isomerase YddE	145	0	-
ydeH_2	Diguanylate cyclase YdeH	146	12	ydeH (0/5), ydeH_1 (146/128)
ydeO_2	transcriptional regulator YdeO	144	0	ydeO (146/135)
yecD_2	Isochorismatase family protein YecD	1	132	yecD (145/122)
yedA_2	putative inner membrane transporter YedA	145	0	yedA (146/134), yedA_1 (12/14)
yedQ	putative diguanylate cyclase YedQ	143	10	yedQ_1 (0/1), yedQ_2 (4/134), yedQ_3 (0/2)
yedQ_2	putative diguanylate cyclase YedQ	4	134	yedQ (143/10), yedQ_1 (0/1), yedQ_3 (0/2)
yedR	Inner membrane protein YedR	146	13	-
ygbF	CRISPR-associated endoribonuclease Cas2	146	13	-
ygbT_1	CRISPR-associated endonuclease Cas1	146	0	ygbT_2 (145/13)
ygbT_2	CRISPR-associated endonuclease Cas1	145	13	ygbT_1 (146/0)
ygcS_1	Inner membrane metabolite transport protein YgcS	146	13	ygcS (0/2), ygcS_2 (146/123), ygcS_3 (145/103), ygcS_4 (146/0)
ygcS_4	Inner membrane metabolite transport protein YgcS	146	0	ygcS (0/2), ygcS_1 (146/13), ygcS_2 (146/123), ygcS_3 (145/103)
ygeX	Diaminopropionate ammonia-lyase	146	13	-
yggF	Fructose-1,6-bisphosphatase 2 class 2	0	134	-
ygjH_2	tRNA-binding protein YgjH	7	129	ygiH (137/12), ygjH_1 (2/84)
ygjI_3	Inner membrane transporter YgjI	146	0	ygiI (12/95), ygiI_1 (133/14), ygiI_2 (146/22), ygiI_4 (0/22)
yhaI	Inner membrane protein YhaI	145	13	-

yhbU_1	Peptidase family U32	145	10	yhbU_2 (146/132), yhbU_3 (146/134), yhbU_4 (0/1)
yhdN	General stress protein 69	139	19	yhdN_2 (0/120), yhdN_3 (4/115)
yiaO_2	2,3-diketo-L-gulonate-binding periplasmic protein YiaO precursor	146	0	yiaO_1 (144/107)
yicJ	Inner membrane symporter YicJ	143	12	yicJ_1 (0/122), yicJ_2 (0/56)
yjcD	Putative permease YjcD	18	133	yjcD_1 (145/19), yjcD_2 (128/2), yjcD_3 (0/1)
yjcD_1	Putative permease YjcD	145	19	yjcD (18/133), yjcD_2 (128/2), yjcD_3 (0/1)
yjdF	Inner membrane protein YjdF	0	132	-
yjdL	putative dipeptide and tripeptide permease YjdL	146	0	-
yjiE	HTH-type transcriptional regulator YjiE	0	133	yjiE_1 (1/78)
yjiG	Inner membrane protein YjiG	0	132	-
yjmD	putative zinc-type alcohol dehydrogenase-like protein YjmD	146	3	-
ymfD	Bacillibactin exporter	146	13	-
ynaI_1	Low conductance mechanosensitive channel YnaI	144	15	ynaI (0/119), ynaI_2 (86/12)
ynjC_1	Inner membrane ABC transporter permease protein YnjC	138	2	ynjC_2 (133/123), ynjC_3 (142/134)
ynjF_1	Inner membrane protein YnjF	146	1	ynjF (2/61), ynjF_2 (144/71)
yqiJ	Inner membrane protein YqiJ	146	19	yqiJ_1 (19/1)
ytbE	putative oxidoreductase YtbE	146	1	-
yvqK	Cob(I)yrinic acid a,c-diamide adenosyltransferase	146	0	-
zraR_1	Transcriptional regulatory protein ZraR	146	0	zraR (145/134), zraR_2 (1/0)
zraS_1	Sensor protein ZraS	146	0	zraS (140/134), zraS_2 (6/3)

Supplementary table 5.2. Homologous gene clusters (HGCs) that were found in significantly higher in either *S. sonnei* or *S. flexneri* pan genomes. Also shown are other HGCs given the same gene name annotation from the BLASTp hit, likely due to different variants of the same gene.

Supplementary material – vcfProcess

vcfProcess is a function written in *R* v3.1.2 that takes VCF files (.vcf) from variant caller software, including Samtools and GATK, and processes these files for use with population genetics tools, primarily for use with haploid organisms. This function can filter indels (insertions/deletions), variants in low coverage regions, variants with low genotype quality, and variants problematic repeat regions. The output files will be a concatenated sequence of high likelihood SNPs in the common FASTA format for downstream analysis such as phylogenetic programs, as well as comma delimited files that can be read in Microsoft Excel.

R Script:

```
vcfProcess<-
function(inputfile,repeatfile.present=FALSE,repeatfile,outputfile,DP_low=5,lowqual=20,low.qua
lity.file=TRUE,repeat.output.file=TRUE,indel.file=TRUE,var.SNPs.only=FALSE){

  ######################################   VCF FILE POST PROCESSING
##################################

  if (!require(stringr)){
    install.packages("stringr")
    library(stringr)
  }
  if (!require(seqinr)){
    install.packages("seqinr")
    library(seqinr)
  }

  ######READ IN VCF FILE, REPEAT REGION FILE AND RETAIN
HEADER#################################################################

  input<-read.table(inputfile)
  header_input<-as.matrix(read.table(inputfile,comment.char=" ",sep="\n"))
  end_head<-which(grepl("#CHROM",header_input)==TRUE)
  header<-as.data.frame(header_input[1:end_head])
  names<-unlist(strsplit(header_input[end_head],"\t"))
  format<-which(names=='FORMAT')
  names<-names[10:length(names)]
  rm(header_input,end_head)

  #######SPLIT GT:PL:DP:GQ... SECTION AND MAKE NEW MATRICES OF THESE
VALUES#################################

  geno<-input[,10:ncol(input)]
  s<-unlist(str_split(geno[1,1], ":"))
  no_of_columns<-length(s)
  GT<-matrix(0,nrow(geno),ncol(geno)*no_of_columns)
  l<-seq(1,ncol(GT),no_of_columns)
  for (i in 1:nrow(geno)){
    for (j in 1:ncol(geno)){
      s<-unlist(str_split(geno[i,j], ":"))
      for (k in 1:no_of_columns){
```

```
      GT[i,l[j]+k-1]<-s[k]
    }
   }
  }

  sep_geno<-cbind(input[,1:9],GT)



  ###################REMOVE UNWANTED SNPS
#####################################################################

  ######### REMOVE LOW QUAL SNPS

  qual<-as.integer(sep_geno[,6])
  b<-qual < lowqual
  d<-which(b=='TRUE')
  lowqual<-sep_geno[d,]
  c<-which(b == 'FALSE')
  sep_geno<-sep_geno[c,]
  vcf<-input[c,]
  if (low.quality.file==TRUE){
   if (nrow(lowqual)!=0){
    write.csv(lowqual,file="lowqualsnps.csv")
   } else {print("No low quality SNPs")
   }
  }

  ######## MAKE INDEL FILE (if applicable)

  desc<-sep_geno[,8]
  new_desc<-grepl('INDEL',desc)
  indel<-which(new_desc=='TRUE')
  indels<-sep_geno[indel,]
  if (indel.file==TRUE){
   if (nrow(indels)!=0){
    write.csv(indels,file="indels.csv")
   } else {print("No indels")
   }
  }
  f<-which(new_desc=='FALSE')
  sep_geno<-sep_geno[f,]
  vcf<-vcf[f,]

  ####### REMOVE REPEAT AND MOBILE ELEMENT REGIONS

  if (repeatfile.present==TRUE){
   rep<-as.data.frame(read.csv(repeatfile,header=TRUE))
   pos<-sep_geno[,2]
   res=vector()
   for (i in 1:nrow(rep)){
    y<-c(rep[i,1]:rep[i,2])
    res<-c(res,y)
   }
   q<-is.element(pos,res)
   a<-which(q == 'FALSE')
   sep_geno<-sep_geno[a,]
   vcf<-vcf[a,]
   b<-which(q=='TRUE')
   rep<-sep_geno[b,]
```

```
  if (repeat.output.file==TRUE){
    if (nrow(rep!=0)){
      write.csv(rep, file="SNPsinrepregions.csv")
    } else {print("No repeat or mobile element SNPs")
    }
  }
}


#### DETERMINE ORDER OF FORMAT FIELDS ####

s<-unlist(str_split(vcf[1,format], ":"))
GT<-which(s=='GT')
DP<-which(s=='DP')
PL<-which(s=='PL')
DP4<-which(s=='DP4')

####### SPLIT DATA INTO 1,2 AND 3 ALT ALLELES
#################################################################

alt<-sep_geno[,5]
b<-which(grepl(",",alt)==TRUE)
sep_geno2<-sep_geno[b,]
vcf2<-vcf[b,]
a<-which(grepl(",",alt)==FALSE)
sep_geno<-sep_geno[a,]
vcf<-vcf[a,]

####### 1 ALT ALLELE ASSIGNING
#############################################################################

ref<-sep_geno[,4]
alt<-sep_geno[,5]
geno<-sep_geno[,seq(GT+9,ncol(sep_geno),no_of_columns)]
genotype<-as.matrix(cbind(ref,alt,geno))

output=matrix(0,nrow(genotype),ncol(genotype))
rownames(output)=rownames(genotype)
colnames(output)=colnames(genotype)

####### ASSIGN NUCLEOTIDE TO GENOTYPE

for (i in 1:nrow(genotype)){
  for (j in 3:ncol(genotype)){
    if (genotype[i,j]=='0/0'){
      output[i,j]=genotype[i,1]}
    else  {if (genotype[i,j]=='1/1'){
      output[i,j]=genotype[i,2]
    }
      else {
        output[i,j]='N'}
    }
  }
}

output<-output[,3:ncol(output)]

######## low read for all mark as '?'
if (length(DP)!=0){
  read<-data.matrix(as.matrix(sep_geno[,seq(DP+9,ncol(sep_geno),no_of_columns)]))
```

```
  class(read)<-"numeric"
  snprd<-read<DP_low
  lowread<-which(snprd == 'TRUE')
  output[lowread]<-'?'
 } else if (length(DP4)!=0){
  read<-data.matrix(as.matrix(sep_geno[,seq(DP4+9,ncol(sep_geno),no_of_columns)]))
  read2<-matrix(0,nrow(read),ncol(read))
  for (i in 1:nrow(read)){
    for (j in 1:ncol(read)){
      read2[i,j]<-sum(as.numeric(unlist(str_split(read[i,j], ","))))
    }
  }
  class(read2)<-"numeric"
  snprd<-read2<DP_low
  lowread<-which(snprd == 'TRUE')
  output[lowread]<-'?'
 }

 ############REMOVE snps that have no alt allele (may have been included due to hetero call)

 alt1<-as.vector(alt)
 new21<-as.matrix(cbind(alt1,output))
 new22<-matrix(0,nrow(new21),ncol(new21))

 for (i in 1:nrow(new21)){
  for (j in 2:ncol(new21)){
    if (new21[i,j]==new21[i,1]){
      new22[i,j]='FALSE'}
    else  {if (new21[i,j]!=new21[i,1]){
      new22[i,j]='TRUE'
    }
    }
  }
 }

 new23<-new22[,2:ncol(new22)]

 d<- lapply(1:nrow(new23), function(i){
  all(as.logical(new23[i,]))
 }
 )
 g<-do.call(rbind,d)

 alt_present<-which(g=='FALSE')
 output<-output[alt_present,]
 vcf<-vcf[alt_present,]

 ################ 2nd AND 3rd ALLELE ASSIGNING
############################################################

 if (nrow(sep_geno2)!=0){

  ###SPLIT 2 and 3 ALT ALLELE DATA

  twoalt<-which(nchar(as.matrix(sep_geno2[,5]))!=5)
  threealt<-which(nchar(as.matrix(sep_geno2[,5]))==5)
  sep_geno3<-sep_geno2[threealt,]
  sep_geno2<-sep_geno2[twoalt,]
  vcf3<-vcf2[threealt,]
  vcf2<-vcf2[twoalt,]
```

```
####2 ALT ALLELE ASSIGNING

ref<-sep_geno2[,4]
alt_split<-t(as.data.frame(strsplit(as.character(sep_geno2[,5]),",")))
alt1<-as.vector(alt_split[,1])
alt2<-as.vector(alt_split[,2])

geno2<-sep_geno2[,seq(PL+9,ncol(sep_geno2),no_of_columns)]
genotype2<-as.matrix(cbind(ref,alt1,alt2,geno2))

output2=matrix(0,nrow(genotype2),ncol(genotype2))
rownames(output2)=rownames(genotype2)
colnames(output2)=colnames(genotype2)

for (i in 1:nrow(genotype2)){
  for (j in 4:ncol(genotype2)){
    s<-unlist(str_split(genotype2[i,j], ","))
    if (max(s)==min(s)){
      output2[i,j]='N'}
    else if (length(which(s==min(s))==1)==TRUE){
      if (which(s==min(s))==3){
        output2[i,j]=genotype2[i,2]
      } else if (which(s==min(s))==6){
        output2[i,j]=genotype2[i,3]
      } else if (which(s==min(s))==1|which(s==min(s))==4){
        output2[i,j]=genotype2[i,1]
      } else {
        output2[i,j]='N'
      }
    } else {
      output2[i,j]='N'
    }
  }
}

output2<-as.matrix(output2[,4:ncol(output2)])
if (ncol(output2)==1){
  output2<-t(output2)
}

######## low read for all mark as '?'

if (length(DP)!=0){
  read<-data.matrix(as.matrix(sep_geno2[,seq(DP+9,ncol(sep_geno2),no_of_columns)]))
  class(read)<-"numeric"
  snprd<-read<DP_low
  lowread<-which(snprd == 'TRUE')
  output2[lowread]<-'?'
} else if (length(DP4)!=0){
  read<-data.matrix(as.matrix(sep_geno2[,seq(DP4+9,ncol(sep_geno2),no_of_columns)]))
  read2<-matrix(0,nrow(read),ncol(read))
  for (i in 1:nrow(read)){
    for (j in 1:ncol(read)){
      read2[i,j]<-sum(as.numeric(unlist(str_split(read[i,j], ","))))
    }
  }
  class(read2)<-"numeric"
  snprd<-read2<DP_low
  lowread<-which(snprd == 'TRUE')
```

```
    output2[lowread]<-'?'
  }

  ############REMOVE snps that have no alt allele (may have been included due to hetero
call)

  new21<-as.matrix(cbind(alt1,alt2,output2))
  new22<-matrix(0,nrow(new21),ncol(new21))

  for (i in 1:nrow(new21)){
    for (j in 3:ncol(new21)){
      if (new21[i,j]==new21[i,1]|new21[i,j]==new21[i,2]){
        new22[i,j]='FALSE'}
      else  {
        new22[i,j]='TRUE'
      }
    }
  }

  new22<-as.matrix(new22[,3:ncol(new22)])
  if (ncol(new22)==1){
    new22<-t(new22)
  }

  d<- lapply(1:nrow(new22), function(i){
    all(as.logical(new22[i,]))
  }
  )
  g<-do.call(rbind,d)

  alt_present<-which(g=='FALSE')
  output2<-as.matrix(output2[alt_present,])
  if (ncol(output2)==1){
    output2<-t(output2)
  }
  vcf2<-vcf2[alt_present,]

  ########## THREE ALT ASSIGN ##########

  if (nrow(sep_geno3)!=0){

    ref<-sep_geno3[,4]
    alt_split<-t(as.data.frame(strsplit(as.character(sep_geno3[,5]),",")))
    alt1<-as.vector(alt_split[,1])
    alt2<-as.vector(alt_split[,2])
    alt3<-as.vector(alt_split[,3])

    geno3<-sep_geno3[,seq(PL+9,ncol(sep_geno3),no_of_columns)]
    genotype3<-as.matrix(cbind(ref,alt1,alt2,alt3,geno3))

    output3=matrix(0,nrow(genotype3),ncol(genotype3))
    rownames(output3)=rownames(genotype3)
    colnames(output3)=colnames(genotype3)

    for (i in 1:nrow(genotype3)){
      for (j in 4:ncol(genotype3)){
        s<-unlist(str_split(genotype3[i,j], ","))
        if (max(s)==min(s)){
          output3[i,j]='N'}
        else if (length(which(s==min(s))==1)==TRUE){
```

```
      if (which(s==min(s))==3){
        output3[i,j]=genotype3[i,2]
      } else if (which(s==min(s))==6){
        output3[i,j]=genotype3[i,3]
      } else if (which(s==min(s))==9|which(s==min(s))==10){
        output3[i,j]=genotype3[i,4]
      } else if (which(s==min(s))==1|which(s==min(s))==4|which(s==min(s))==7){
        output3[i,j]=genotype3[i,1]
      } else {
        output3[i,j]='N'
      }
    } else {
      output3[i,j]='N'
    }
  }
}

output3<-as.matrix(output3[,5:ncol(output3)])
if (ncol(output3)==1){
  output3<-t(output3)
}

######## low read for all mark as '?'

if (length(DP)!=0){
  read<-data.matrix(as.matrix(sep_geno3[,seq(DP+9,ncol(sep_geno3),no_of_columns)]))
  class(read)<-"numeric"
  snprd<-read<DP_low
  lowread<-which(snprd == 'TRUE')
  output3[lowread]<-'?'
} else if (length(DP4)!=0){
  read<-data.matrix(as.matrix(sep_geno3[,seq(DP4+9,ncol(sep_geno3),no_of_columns)]))
  read2<-matrix(0,nrow(read),ncol(read))
  for (i in 1:nrow(read)){
    for (j in 1:ncol(read)){
      read2[i,j]<-sum(as.numeric(unlist(str_split(read[i,j], ","))))
    }
  }
  class(read2)<-"numeric"
  snprd<-read2<DP_low
  lowread<-which(snprd == 'TRUE')
  output3[lowread]<-'?'
}

############REMOVE snps that have no alt allele (may have been included due to hetero
call)

new21<-as.matrix(cbind(alt1,alt2,alt3,output3))
new22<-matrix(0,nrow(new21),ncol(new21))

for (i in 1:nrow(new21)){
  for (j in 4:ncol(new21)){
    if (new21[i,j]==new21[i,1]|new21[i,j]==new21[i,2]|new21[i,j]==new21[i,3]){
      new22[i,j]='FALSE'}
    else  {
      new22[i,j]='TRUE'
    }
  }
}
```

```
    new22<-as.matrix(new22[,4:ncol(new22)])
    if (ncol(new22)==1){
      new22<-t(new22)
    }

    d<- lapply(1:nrow(new22), function(i){
      all(as.logical(new22[i,]))
    }
    )
    g<-do.call(rbind,d)

    alt_present<-which(g=='FALSE')
    output3<-as.matrix(output3[alt_present,])
    if (ncol(output3)==1){
      output3<-t(output3)
    }
    vcf3<-vcf3[alt_present,]

    ###### combine output2&3 and vcf2&3

    pos2<-as.numeric(vcf2[,2])
    pos3<-as.numeric(vcf3[,2])
    output2<-cbind(pos2,output2)
    output3<-cbind(pos3,output3)
    output2<-rbind(output2,output3)
    output2<-output2[order(as.numeric(output2[,1])),]
    vcf2<-rbind(vcf2,vcf3)
    vcf2<-vcf2[order(as.numeric(vcf2[,2])),]
    output2<-output2[,2:ncol(output2)]

  }

  ###### combine output&output2 and vcf&vcf2

  pos<-as.numeric(vcf[,2])
  pos2<-as.numeric(vcf2[,2])
  output<-cbind(pos,output)
  output2<-cbind(pos2,output2)
  output<-rbind(output,output2)
  output<-output[order(as.numeric(output[,1])),]
  output<-output[,2:ncol(output)]
  vcf<-rbind(vcf,vcf2)
  vcf<-vcf[order(as.numeric(vcf[,2])),]
}

 ################ Remove SNPS that are shared by all isolates (i.e. no variation in isolates,
only from reference)

 if (var.SNPs.only==TRUE){
  ref<-as.character(vcf[,4])
  mat<-cbind(ref,output)
  new22<-matrix(0,nrow(mat),ncol(mat))

  for (i in 1:nrow(mat)){
    for (j in 2:ncol(mat)){
      if (mat[i,j]==mat[i,1]){
        new22[i,j]='FALSE'}
      else  {
        new22[i,j]='TRUE'
      }
```

```
    }
   }

   new22<-new22[,2:ncol(new22)]

   d<- lapply(1:nrow(new22), function(i){
    all(as.logical(new22[i,]))
   }
   )
   g<-do.call(rbind,d)

   var_present<-which(g=='FALSE')
   output<-output[var_present,]
   vcf<-vcf[var_present,]
  }

  ################## writing different file
types################################################

  ########## Write VCF file of processed SNPs

  v<-as.data.frame(apply(vcf, 1, paste, collapse="\t"))
  rownames(v)<-NULL
  names(header)<-names(v)
  u<-rbind(header,v)
  write.table(u, file= paste(outputfile,".vcf",sep = ""),row.names
=FALSE,sep="\t",quote=FALSE,col.names=FALSE)

  ########### CSV file with SNP genotypes only with reference sequence and SNP position

  colnames(output)<-names
  Reference<-as.character(vcf[,4])# makes single row of ref alleles (from reference sequence)
  Position<-vcf[,2]
  forcsv<-cbind(Position,Reference,output)
  write.csv(forcsv,file=paste(outputfile,".csv",sep = ""),row.names=FALSE)

  ########### Fasta file of all isolates (without ref)
  output_fast<-t(output)
  forfasta<-as.list(apply(output_fast, 1, paste, collapse=" "))
  write.fasta(forfasta,names,nbchar=60,file.out=paste(outputfile,".fasta",sep = ""),open="w")

  ########### Fasta file of all isolates (with ref)

  namesfasta<-c("Reference",names)
  forfastaref<-rbind(Reference,output_fast)
  forfastaref<-as.list(apply(forfastaref, 1, paste, collapse=" "))
  write.fasta(forfastaref,namesfasta,nbchar=60,file.out=paste(outputfile,"_with_ref.fasta",sep =
""),open="w")}
```