The genomic epidemiology of shigellosis in sub-Saharan Africa

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

George E. Stenhouse

August 2022

Acknowledgements

I would like to thank my primary supervisor, Kate Baker, for her consistent support and guidance. I would not be the scientist I am today without her encouragement. Thank you so much for reading and re-reading all the drafts of this thesis, regardless of how rough they were. Your advice has been infinitely helpful and highly appreciated. I would also like to thank my secondary supervisor, Miren Iturriza-Gómara, whose advice and encouragement have greatly improved my PhD project and ultimately this thesis. Thank you for pushing me to do my best.

This thesis would not have been possible without the aid of Becky Bengtsson, who taught me bioinformatics basics and answered my many, many questions. Thank you for always being there to discuss bioinformatic methods and for encouraging me when I didn't feel like I was making progress. Thank you, also, to Caisey Pulford, who has been a great bioinformatic support, and to the rest of the Bakery, who have all helped me grow as a scientist and kept me motivated.

Many thanks to all my collaborators who made this work possible. Particular thanks to Karen Keddy, Juno Thomas, and Anthony Smith whose expertise and advice have been invaluable to this project.

My sincere thanks to my other PhD candidate friends, Lewis Fischer, Eleri Ashworth, and Fredi Langendonk, who have regularly taken the time to listen, you have all helped keep me sane. I would also like to thank my family for their constant love and support. Thank you for always believing in me.

Abstract

Shigellosis is a leading cause of diarrhoeal death globally, though the highest prevalence is in low- and middle-income countries with the greatest disease burden falling on children under five years of age. Increasing antimicrobial resistance is a growing hindrance to ongoing effective treatment. Whole genome sequence analysis (WGSA) of the aetiological bacterial genus, *Shigella*, has been effectively used to better understand the pathogen and disease epidemiology, and antimicrobial resistance (AMR). Shigellae from sub-Saharan Africa has been relatively understudied using WGSA. This thesis reports the findings of three studies applying WGSA to sub-Saharan African *Shigella* at different population levels (hospital, national and sub-continent). This thesis finds, for the first time, a link between the HIV pandemic and *Shigella* evolution, with endemic strain diversification and successful introductions coinciding with the HIV epidemic in South Africa. While widespread MDR was confirmed across sub-Saharan Africa, an emergent, pan-susceptible *S. flexneri* 2a lineage was also identified in South Africa. The greater drug susceptibility of this lineage was found to be linked to the absence of the known MDR element the *Shigella* resistance locus (SRL). I also found new evidence to suggest that retention of the large virulence plasmid (pINV) is a hinderance to strain success in *S. sonnei*, which was found to vary between the identified South African sub-Clades. The results from the national study also confirm the ability for multiple *S. flexneri* strains to coexist while supporting the inability of *S. sonnei* strains to do likewise. Vaccine development is a current strategy to reduce shigellosis cases in the face of increasing AMR. Previously undetected strains from both serotypes were described in this thesis, highlighting the need for the global application of WGSA for *Shigella*. This work provides key insights for the tracking of AMR emergence and spread, and effective *Shigella* vaccine development and future deployment, issues critical for ongoing effective treatment of shigellosis both within sub-Saharan Africa and across the globe. The results also highlight the importance of considering disease dynamics in HIV+ populations and serotype transmission pathway preference when developing public healthcare policy aimed at reducing shigellosis.

Abbreviations

AA	Amino acid
AMR	Antimicrobial resistance
AMP	Ampicillin
BAPS	Bayesian Analysis of Population Structure
BEAST	Bayesian Evolutionary Analysis by Sampling Trees software
CGR	The Centre for Genomic Research, University of Liverpool
CHL	Chloramphenicol
CRO	Ceftriaxone
EIEC	Enteroinvasive Escherichia coli
FQR	Fluoroquinolone resistance
GBD	Global Burden of Disease
GEMS	The Global Enteric Multicentre Study
GERMS-SA	The Group for Enteric, Respiratory and Meningeal Diseases Surveillance in South Africa
LMIC	Low- and middle-income countries
MASCOT	Marginal approximation of the structured coalescent
MDR	Multidrug resistant/resistance
MDRE	Multidrug resistance element
MGE	Mobile genetic elements
MLST	Multi-locus sequence typing
MRCA	most recent common ancestor
MSD	Moderate-to-severe diarrhoea
MSM	Men who have sex with men
NGS	Next generation sequencing

PAI	Pathogenicity Island
pINV	Large virulence plasmid
PG	Phylogroup
QRDR	Quinolone resistance determining region
R0	Basic reproduction number
Sb	*Shigella boydii*
Sf	*Shigella flexneri*
SHI	*Shigella* pathogenicity island
SNP	Single nucleotide polymorphism
SRL	*Shigella* resistance locus
SRL-PAI	*Shigella* resistance locus pathogenicity island
Ss	*Shigella sonnei*
ST	Sequence type
STR	Streptomycin
TA	Toxin-antitoxin
TCN	Tetracycline
TMP-SUL	Cotrimoxazole
WASH	Water, hygiene, and sanitation
WGS	Whole genome sequence/sequencing
WGSA	Whole genome sequence analysis
T3SS	Type III secretion system

Contents

Figures

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Chapter 7

Supplementary

Tables

Chapter 1

Chapter 2

Chapter 3

Chapter 4

Chapter 5

Chapter 6

Supplementary

Chapter 1

Introduction

1.1. An overview of moderate-to-severe diarrhoeal disease in humans

1.1.1. The global burden of diarrhoeal disease

Diarrhoeal diseases are the eleventh leading cause of death globally, according to the Global Burden of Disease (GBD) study in 2019 (https://vizhub.healthdata.org/gbd-results/) [1]. The diarrhoeal disease burden is, however, unevenly distributed. Geographically, the greatest disease burden falls on those in living in low- and middle-income countries (LMIC); diarrhoeal disease is the third leading cause of death in sub-Saharan Africa [1]. Meanwhile, children under the age of five years face the greatest risk of infection, morbidity, and mortality. The highest diarrhoeal mortality rate in children under five was in sub-Saharan Africa in 2017 (204.6 deaths per 100,000 children) [2]. Diarrhoeal deaths in children are in decline; however, the rate of decline is slowest in those under five years of age, this is particularly true in high burden regions such as sub-Saharan Africa and South Asia [3].

An important risk factor for diarrhoeal disease is access to clean water, hygiene, and sanitation (WASH) [4]. Availability of WASH is highly linked to the wealth and industrialisation of a region. Diarrhoeal disease being the second leading cause of death among poorest one billion people in the world shows that diarrhoeal mortality is highly linked to wealth [5]. Those living in urban regions typically have a greater access to WASH provisions than those in rural regions, although some urban regions are also poorly provided for [4, 6, 7].

Open defecation, a practice often necessitated by a lack of access to sanitation, is high in Sub-Saharan Africa; approximately half of all people who practice open defecation are in sub-Saharan Africa [8]. The practice is decreasing in the region, though at a slower rate than in Central and Southern Asia which is likely a contributing factor for the slower declines in infant diarrhoeal mortality rates [6, 8].

While many different pathogens contribute to diarrhoeal disease globally, rotaviral enteritis, caused by rotavirus, is the leading cause of diarrhoeal death, accounting for 15.2% of all-age diarrhoeal deaths and 29.3% of the deaths in children under five years of age in 2015 [9-11]. Shigellosis, caused by *Shigella* bacteria, is the second leading cause, accounting for approximately 6.0-13.2% of diarrhoeal

deaths globally in 2015 [9, 10, 12]. Though in children under the age of five years old, cryptosporidiosis, caused by cryptosporidium parasites, is the second leading cause of diarrhoeal deaths (12.1% of deaths), followed by shigellosis (11.0%) [9-12].

1.1.2. The shigellosis global disease burden

The greatest disease burden of shigellosis falls on children under the age of five, accounting for around one third of shigellosis deaths [10, 11]. Furthermore, repeated infections in children can result in long-term health effects such as stunted growth, cognitive impairment, and chronic, functional bowel disorders, adding to the disease burden [10, 13-19]. Geographically the greatest burden is on those living in sub-Saharan Africa and southern Asia, repeated childhood infection is common in these high prevalence regions [10, 11].

The mainstay of treatment is supportive care; however, hospital cases are typically provided with antimicrobial treatment to minimise mortalities and chronic morbidities [13, 20]. Widespread multidrug resistance, and the continued development of further resistance, are a threat to the ongoing effective treatment of shigellosis and have been deemed important enough that fluoroquinolone resistant (FQR) *Shigella* strains have been included on the WHO priority pathogen list and *Shigella* are a WHO recommended priority for vaccine development [21, 22]. The greatest impact of increasing antimicrobial resistance will fall on those in high shigellosis burden regions as access to healthcare in LMIC countries is typically harder than for those living in high income countries.

1.2. Shigellosis aetiology and pathogenesis

Shigellosis is caused by infection with *Shigella,* a group of Gram-negative bacteria which are highly adapted to humans [23, 24]. Shigellosis has a distinctive, enteroinvasive pathogenesis which is mediated by a type III secretion system (T3SS) and secreted effector proteins, encoded by genes on the large virulence plasmid (pINV) [20, 25, 26].

The *Shigella* bacteria initially gain entry to the colonic mucosa via M-cells, specialised epithelial cells associated with lymphoid tissues of the intestinal mucosa which sample antigens from within the gastrointestinal tract [27]. Antigens captured at the apical side are rapidly transported across the epithelium via transcytosis through the M-cell. Macrophages waiting at the basal surface of the M-cells will phagocytose larger antigens such as bacteria. *Shigella* can stall phagocytosis, escaping the phagocytic vacuole and triggering macrophage necrosis [28]. The released cytokines promote a pro-inflammatory response from the innate immune system.

Macrophage cell death results in extracellular release of the *Shigella* bacteria, which are then able to invade the basal side of the colonic epithelium in a T3SS mediated manner [28]. The *Shigella* bacteria replicate intracellularly and spread between epithelial cells via the connected lateral surfaces. This is aided through the inhibition of epithelial cell death and autophagy. Secreted factors also suppress epithelial cell to innate immune system signalling by inhibiting NFκB, slowing the innate immune response and prolonging infection.

The early inflammation induced by colonic invasion by *Shigella* results in disruption of the epithelial tight junctions allowing further colonic invasion by *Shigella* in the intestinal spaces [28]. Epithelial cell rupture produces colonic lesions and the typical dysentery seen with shigellosis. As the disease progresses, the inflammatory immune response is often effective at eliminating the infection [28]. Though in some cases infection spreads into the bloodstream and becomes systemic. The factors driving an invasive disease presentation are not yet understood.

1.3. *Shigella* evolution and typing nomenclature

1.3.1. *Shigella*: a human adapted pathovar of *Escherichia coli*

Shigella are a group of human-adapted pathogenic bacteria which form a pathovar of *E. coli* (Figure 1.1) [29]. Previous research suggests that the evolution from intra-intestinal ancestral *E. coli* to enteroinvasive and intracellular *Shigella* has likely occurred on multiple occasions [29]. Convergent evolution of human adaptation of the multiple strains resulted in a diverse group of bacteria within the *E. coli* clade sharing a common pathogenesis.

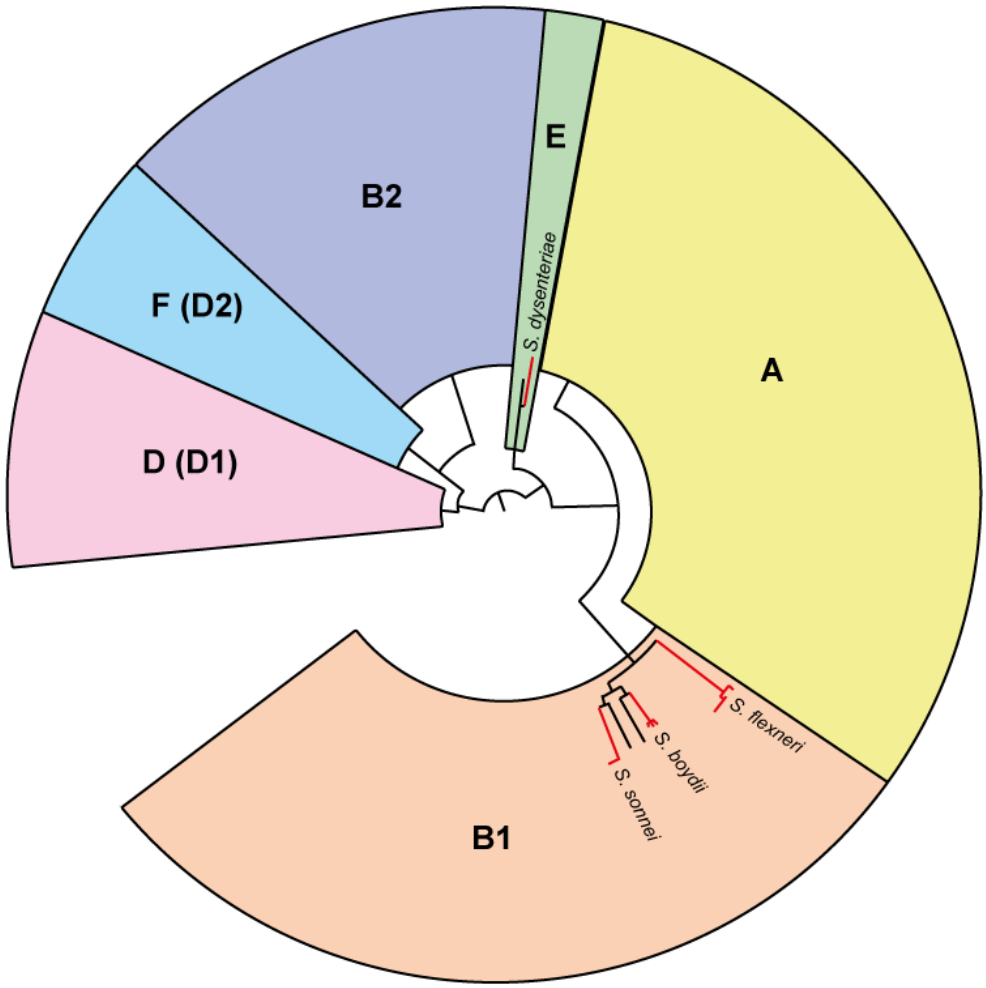

Figure 1.1. Representation of the Shigella/E. coli clade based on a single copy gene ortholog-based maximum likelihood phylogeny showing the dispersal of the Shigella serotypes within the E. coli population.
Phylogroups (A, B1, B2, D, F and E) are shown by background colour. All *Shigella* serogroups (red branches) belong to Phylogroup B1 (peach background) except *S. dysenteriae* which belongs to Phylogroup E (green background). Each serogroup forms a distinct cluster within the *E. coli/Shigella* clade, each evolving from a different ancestral *E. coli* strain.
Schematic based on phylogeny from [29].

This evolution was likely initiated by the acquisition of the pINV which enabled invasion of the colonic mucosa. Subsequent convergent evolution, through the loss of metabolic pathways and the gaining of virulence has resulted in human host specificity, though sporadic isolation from non-human reservoirs has been reported [30]. A similar pathogenesis has been seen in some *E. coli* which also possess a pINV, termed enteroinvasive *E. coli* (EIEC), though human adaptation has not occurred in these strains and are therefore not considered shigellae [23, 24].

1.3.2. *Shigella* population structure and typing nomenclature

A major consideration for vaccine development is genetic diversity, particularly in the targeted surface antigens. The main typing nomenclature for *Shigella,* based on the structure of the surface O-antigen of the cell wall lipopolysaccharide, groups them into four species or serogroups (*Shigella flexneri*, *Shigella sonnei*, *Shigella boydii* and *Shigella dysenteriae*), all containing multiple serotypes except *S. sonnei*.

The *S. sonnei* O-antigen was likely obtained from *Plesimonas shigellosides,* replacing the previous O-antigen [31]. Functional O-antigen genes, highly related to those in *P. shigellosides* are carried on the pINV while loss of function mutations have occurred in most of the chromosomally located O-antigen genes [31]. The stability of this O-antigen variant in *S. sonnei* may in part be due to the expression of an O-antigen capsule; no other *Shigella* serotype is known to have a functional O-antigen locus. The O-antigen capsule has been shown to decrease virulence but promote pathogen persistence in host tissues, likely through immune evasion [32].

Since the use of whole genome sequence analysis (WGSA) for *Shigella*, strains have started to be typed according to their evolutionary relationships to each other. The evolutionary relationships can be modelled with phylogenetics based on the single nucleotide polymorphisms (SNPs) differences between the included isolates. Each serogroup has its own nomenclature based on the modelled global phylogenetic population structure [33-36].

The *S. flexneri* phylotyping nomenclature does not completely overlap with the serotyping nomenclature, likely due to the ability of *S. flexneri* to switch serotype [33]. Population clustering has been used to define seven Phylogroups of *S. flexneri* which was the current extent of the phylotyping scheme in *S. flexneri* at the beginning of this thesis [33]. Another, yet unnamed Phylogroup of *S. flexneri* has also been identified in the literature, a more distantly related Phylogroup compared to the relatedness between the named Phylogroups, which contains exclusively serotype 6 isolates, the only Phylogroup to include serotype 6 [33, 37]. Phylogroups 1 to 7 all contain multiple serotypes.

During this thesis a new typing scheme was developed, based on core genome multi-locus sequence types, which covered the known *Shigella/E. coli* clade [38]. However, as the scheme was not published until 2022 it was not used in this thesis.

The genetic diversity of the known *S. sonnei* population is similar to that of a single Phylogroup of *S. flexneri*, however, a detailed phylotyping nomenclature has been developed for *S. sonnei* [37, 39]. Also based on the population clustering, previously defined into 5 five Lineages [35]. Each Lineage is divided into Clades, which are themselves divided into Sub-Clades [39].

Four Lineages of *S. dysenteriae* have also been defined [34]. This serogroup has been studied less than *S. sonnei* or *S. flexneri* and so a more detailed nomenclature has not been defined. Similarly, *S. boydii* is also relatively understudied as a serogroup, even the population structure is less well defined though three Clades have been identified [36].

1.4. Genomic epidemiology of infectious diseases

1.4.1. Genomics within infectious disease epidemiology

The use of genetics and genome techniques for epidemiology can be helpful in several ways: 1) identifying the aetiological agent, 2) identifying virulence factors and understanding the pathogenesis, 3) identification of sources/reservoirs and transmission pathways, 4) strain identification and tracking, 5) strain transmissibility, and 6) the intercommunity transmission patterns [40].

Culture-based pathogen identification is not always possible due to the inability to culture some pathogens [40]. In this case it can be possible to identify pathogen DNA directly from samples for pathogen identification. Culture free DNA detection helped identify the association between the Kaposi's Sarcoma-associated herpes virus [41, 42]. Initially the isolated DNA was identified as herpes virus DNA, through comparison with known viral sequences. Then a case-control study was able to confirm the identified herpes virus was linked to Kaposi's sarcoma [41, 42].

Similarly, whole genome-based methods are also effective for differentiating between EIEC and *Shigella*, PCR-based methods are insufficient. Genomic comparisons can show if a newly identified strain is more closely related to previously identified *E. coli* or *Shigella*. Culture-based biochemical tests can differentiate between the species, as *Shigella* has different metabolic properties to *E. coli*, however the identification is not 100% in agreement with the genome-based methods.

The genetic comparison of strains with observable differences in virulence can lead to the identification of associated virulence factors [40]. The function of the identified virulence factors can then be studied for a better understanding of the pathogen biology and pathogenesis. Such methods were effectively used to identify novel allelic forms of virulence determinant carrying genomic islands by comparing a community-acquired methicillin-resistant *Staphylococcus aureus* isolate, associated with increased risk of mortality, against less virulent hospital-acquired strains [43].

The comparison of the genetic similarity can also be used to link strains from different sources (i.e., patient and environment), providing support for the transmission of strains between the two sources, identifying likely reservoirs and transmission pathways [40]. The high similarity of *E. coli* isolates from chickens and extraintestinal pathogenic *E. coli* in humans showed that chickens were a reservoir of urinary tract infection causing *E. coli* [44, 45].

The high-definition level of strain identification allows tracking of individual strains or important genetic elements, such as AMR genes, though time and space [40]. This can be important for effective treatment and/or vaccination, for annually updating the flu vaccine for example. Genetic tracking of COVID-19 variants globally also aided in the development of effective public health responses in advance of 'waves' of increased transmission from new variant introductions to a region.

Strain transmissibility can be defined as the basic reproduction number (R_0) which can be is determined through contact tracing. Finding R_0 in this way requires knowledge of the exact transmission chain however, which can be difficult to determine for some infections such as sexually transmitted infections where it is not always clear which partner became infected first [40]. Genetic techniques can be used to determine the transmission chain and therefore R_0. It is also possible to estimate R_0 from a random sample from the study population with whole genome sequence analysis (WGSA) [46-49].

Genetic comparisons combined with patient-related data, such as geographic location, can provide information on intercommunity transmission [40]. Such methods have been used to identify where and when HIV became a human pathogen and the transmission events involved in its global dissemination [50].

1.4.2. Genomic epidemiology techniques

1.4.2.1. *Whole genome sequencing and sequence read quality processing*

There are now many techniques for sequencing genomes. Next-generation sequencing (NGS) techniques are generally defined as those which are capable of high throughput sequencing, involving highly parallel sequencing; NGS encapsulates several generations of sequencing technologies [51]. These HTP technologies produced high accuracy through high sequencing depth and subsequent generation of consensus sequence of multiple reads [51].

Earlier NGS techniques produced short sequence reads (50-300 nucleotides). The short read technologies, however, have poor accuracy across repeat regions and are therefore not often capable of producing complete genome assemblies or accurately determining repeat region alleles [51]. Later NGS have started to produce longer sequence reads to combat these issues, produce faster sequencing at lower cost and detect epigenetic modifications during sequencing.

Processing of raw reads is required before analysis to ensure adequate quality. As the DNA fragments are sequenced, the quality of the read at each base is given a quality score to reflect the likely accuracy of the read. Removal of bases with poor quality scores, and reads with poor overall quality scores, ensures high accuracy of consensus sequences during analysis [52-54]. Certain read qualities, such as GC content, are used to assess read quality and likelihood of contamination so that poor quality and contaminated genomes can be excluded from further analysis [55]. Additionally, removal of adaptor sequences, added during the sequencing process, ensures all included sequences are from the target genome.

1.4.2.2. *Phylogenetics*

When phylogenetics, the modelling of evolutionary relationships, is used to model microbial evolution through vertical inheritance (clonal reproduction) it is typically performed using a selection of genes, common across all individuals whose relationships are to be modelled, or single nucleotide polymorphisms (SNPs) [56-58]. A SNP-based approach enables modelling from data across the entire

genome and removes the influence of gene choice, by the researcher, on the results [56]. It is also important, however, to carefully define SNPs, excluding all which arise due to recombination if modelling vertical inheritance. The choice of which SNPs to exclude can also influence the resultant phylogeny [59].

There are multiple statistical methods for phylogenetic modelling, the most used are 1) maximum likelihood and 2) Bayesian statistics. Maximum likelihood statistics aims to find the phylogenetic model with the greatest likelihood given the genetic input data and usually an explicit nucleotide transformation model [60]. Iteration through all the possible models to find the most likely is too computationally taxing, thus a range of different model sampling algorithms have been developed to efficiently sample tree space to find the most likely [61, 62]. The accuracy of the generated topology is evaluated by bootstrapping the input data [63, 64].

Bayesian phylogenetics can incorporate influencing factors beyond those of a maximum-likelihood model, such as sampling date, enabling the inference of a wider range of variables [65]. Bayesian statistics aims to estimate the probability of the model being true for the population of study from which the included individuals were selected (assuming random selection and complete mixing of the study population) [66]. Meanwhile maximum likelihood aims to find the most likely model for the input sequence alignment, with bootstrapping to estimate the model accuracy [60, 64]. The benefits of using Bayesian statistics, however, require the inclusion of informative prior information into the model according to Bayes' theorem.

1.5. Genomic epidemiology of shigellosis

1.5.1. Application of genomic epidemiology methods for shigellosis

The application of WGSA to *Shigella* has already taught us a lot about the genomic epidemiology of shigellosis. The use of phylogenetics has provided strong support for the convergent evolution of multiple *E. coli* ancestor strains into the modern *Shigella* pathovar [67, 68]. While serogroup specific phylogenies, including isolates collected from around the world and across time, has enabled the construction of a global serogroup population structures [33-36].

Using Bayesian statistics, the current global phylogeny dates the emergence of *S. flexneri* phylogroups to between the 1100's and the 1800's [33]. Since emerging, the serogroup has diversified greatly into seven named Phylogroups, to date; another distantly related serotype 6 specific Phylogroup has also been identified in the literature [33, 37]. Each named Phylogroup is internationally dispersed and contain multiple serotypes. Most serotypes also belong to multiple Phylogroups due to serotype switching mediated via phages [69-71].

Dating of *S. sonnei* emergence with similar Bayesian phylogenetic methods estimated the serogroup likely emerged more recently, during the 1600's in Europe [35]. The phylogeny also showed that there are five Lineages, though one is globally dominant (Lineage III) (Figure 1.5). This dominant Lineage has disseminated internationally from Europe, the other less internationally successful Lineages are predominantly found in Europe. Due to the global dominance of a single Lineage, *S. sonnei* strains across the global are more closely related compared to the diversity seen in *S. flexneri* [33, 35].

The determination of the genotypic AMR profiles has been used to examine the generalised AMR profiles of the identified Lineages and Phylogroups as well as to track individual AMR determinants. On the global population level, AMR genotyping has confirmed MDR is, globally, widespread in both *S. flexneri* and *S. sonnei* [33, 35]. *In silico* AMR genotyping, alongside phylogenetics, was effectively used to identify and track the horizontal transfer of an azithromycin resistance conferring plasmid (pKSR100) in Europe, showing that the acquisition of this plasmid was driving epidemics in the region

[72]. Meanwhile, phylogeography has been used on a national level to model the introduction and spread of AMR *S. sonnei* in Vietnam which demonstrated initial establishment in Ho Chi Minh city followed by multiple introductions to Khanh Hoa and the subsequent, successful introduction to Hue [73].

1.5.2. Global epidemiology of shigellosis

Each *Shigella* serogroup has a distinct epidemiology. Of the four, *S. flexneri* and *S. sonnei* are the most prevalent globally. Both are found across the world with endemic transmission being more associated with LMIC and epidemics and travel-related cases with high-income countries, though more recently endemic strains have been circulating in the community of men who have sex with men (MSM) across the world regardless of country wealth. Serogroup dominance has been observed as being associated with the level of industrialisation of a region, with a switch from *S. flexneri* dominance to *S. sonnei* dominance as a country industrialises [74].

There are several theories for why this serogroup association with level of industrialisation exists. The two most likely, and non-mutually exclusive, theories suggest that people in LMIC are more exposed than those in high-income countries to *Plesiomonas shigelloides* which has a similar O-antigen structure to *S. sonnei* and may provide protective cross immunity to *S. sonnei* specifically WHO. Additionally, the strength of the influence of AMR acquisition and the instability of the pINV in *S. sonnei* suggests a transmission pathway with minimal environmental exposure, while *S. flexneri* transmission potentially involves environmental passage and is therefore transmitted more regularly in places with a poorer access to WASH [24, 76, 77].

The other two serogroups, *S. boydii* and *S. dysenteriae*, are less prevalent though have also been detected around the world. To date no estimate for the emergence of *S. boydii* has been made, likely in part due to the low prevalence of the serogroup globally [36]. However, *S. dysenteriae* Type 1 likely also emerged in Europe, though slightly later than *S. sonnei*, in the 1700's which was a period of industrialisation and increasing population sizes [34]. Strains likely spread across Europe and to the

middle East during the 1800's. From there, in the later part of the century, strains spread to central America and then Across to West Africa. Spread down into Eastern Africa and across to East Asia, from the middle East, likely occurred in the in the 1900's. Many of the strains were identified in the study were from the World Wars and other conflicts or Natural disasters across the world, while at least one international transmission event has been linked to European colonisation. Overall, *S. dysenteriae* appears to be associated with poor conditions.

The limited data on *S. dysenteriae* and *S. boydii* mean that local population dynamics of these serogroups are unknown. The *S. sonnei* phylogeny, however, suggests that the introduction of a new strain leads to clonal replacement, with old strains being replaced by the new strain, driven by the acquisition of AMR [35]. While the *S. flexneri* phylogeny suggests that older strains persist and coexist with newly introduced strains [33].

1.5.3. The role of antimicrobial resistance in *Shigella* epidemiology

The acquisition of AMR typically plays an important role in *Shigella* strain success. Multidrug resistance (MDR) is widespread in both *S. sonnei* and *S. flexneri*. Susceptibility to third generation cephalosporins and fluoroquinolones are still, generally, widespread, though fluoroquinolone resistance (FQR) is now prevalent in Asia and there is evidence of these strains becoming disseminated around the world [78, 79].

Most AMR in *Shigella* is conferred through the acquisition of resistance genes, however FQR can be conferred by the presence of a resistance gene or through *de novo* point mutations in the quinolone resistance determining region (QRDR) [80]. Two point mutations in the *gryA* gene are required for full FQR, a single point mutation results in resistance to quinolones. A third mutation, in the *parC* gene, ameliorates the fitness cost of the double *gryA* mutation and is required for fixation of QRDR point mutation mediated FQR in the strain [78-80].

The acquisition of AMR has been linked to strain success in *S. sonnei* [35]. The success of the globally dominant Lineage III has been attributed to the acquisition of a Tn7/In2 transposable element encoding AMR genes (*dfrA1, sat2* and *aadA1)* conferring resistance to trimethoprim, streptothricin, and aminoglycosides, respectively [35]. Approximately half of the globally successful Lineage III isolates also carried an MDR plasmid (spA), encoding *tetA, strAB,* and *sul2*, and conferring resistance to tetracycline, streptomycin and sulfonamides, respectively [35].

Acquisition of AMR has not been linked to *S. flexneri* success in the same way as for *S. sonnei* as strain appear to coexist rather than replace each other, though a study of *Shigella* in South East Asia did see a clonal replacement driven by AMR in *S. flexneri* on the serotype level [33, 81]. The widespread AMR in *S. flexneri* does, however, point to AMR playing a role in strain success [33]. Another exception to a non-direct link between AMR acquisition and *S. flexneri* strain success has been observed when the strains are sexually transmitted. Both *S. sonnei* and *S. flexneri* strains have been observed as transmitting in this way, observed within the community of men who have sex with men, with epidemics driven by the acquisition of AMR [72].

The MDR conferring *Shigella* resistance locus (SRL) pathogenicity island (PAI) has been acquired by *S. dysenteriae* Lineages at least four times and was the primary source of MDR in the global phylogenetic study [34]. The SRL-PAI, first identified in *S. flexneri*, encodes a multi-drug resistance element (MDRE) which has four AMR genes, *blaOXA-1*, *catA1, aadA1 and, tetA(B),* conferring resistance to ampicillin, chloramphenicol, streptomycin, and tetracycline, respectively [82, 83].

1.6. Epidemiology of shigellosis in sub-Saharan Africa

A review of the literature of all sub-Saharan African *Shigella* papers in the Scopus database, published from 2015 to 2022 and written in English showed that despite being a region of high mortality and morbidity from shigellosis, shigellae from sub-Saharan Africa have been relatively understudied, particularly using WGSA (Table 1.1). Literature review included all papers with *Shigella* or shigellosis, and Africa or Algeria, Angola, Benin, Botswana, Burkina Faso, Burundi, Cameroon, Cabo Verde, Central African Republic, Chad, Comoros, Côte d'Ivoire, Democratic Republic of the Congo, Djibouti, Egypt, Equatorial Guinea, Eritrea, Eswatini, Ethiopia, Gabon, Gambia, Ghana, Guinea, Guinea-Bissau, Kenya, Lesotho, Liberia, Libya, Madagascar, Malawi, Mali, Mauritania, Mauritius, Morocco, Mozambique, Namibia, Niger, Nigeria, Republic of the Congo, Rwanda, Sao Tome Principe, Senegal, Seychelles, Sierra Leone, Somalia, South Africa, South Sudan, Sudan, Tanzania, Togo, Tunisia, Uganda, Zambia, or Zimbabwe, in the title or abstract.

Much of what is known of sub-Saharan African *Shigella* epidemiology comes from small local cross-sectional studies. Most studies are on the aetiological agents of moderate-to-severe diarrhoea (MSD) and typically involve children under the age of five admitted to a single hospital. Although the most common study conducted varies by country (Table 1.1). There is, generally, an under-representation of adults and asymptomatic carriers in the literature (Table 1.1). Much of the literature does not include serotyping of *Shigella* isolates (Table 1.2), with several also not distinguishing between *Shigella* and EIEC, particularly in Western sub-Saharan Africa (Table 1.1). Distinguishing between EIEC and Shigella without WGSA can be difficult; *Shigella* are regularly detected with PCR and primers for identifying the presence of the *ipaH* gene, which is carried by both EIEC and *Shigella* [84].

Cross-sectional studies of the aetiological agents of MSD are important for providing insight into the relative contribution of pathogens to diarrhoeal disease, however, they do not provide an overall estimate for the prevalence of *Shigella*. For this, case-control studies provide a better estimate by measuring asymptomatic carriage, known to sometimes occur with *Shigella* (Table 1.1) [85]. Few

recent studies were identified examining the level of asymptomatic carriage of *Shigella* in sub-Saharan Africa, though cross-sectional studies of asymptomatic food handlers have been conducted in some countries, particularly Ethiopia (Table 1.1). Cross-sectional studies of asymptomatic carriage in combination with MSD cross-sectional studies can provide a more complete picture of *Shigella* prevalence, though the total number of diarrhoea cases must also be known and is generally unknown.

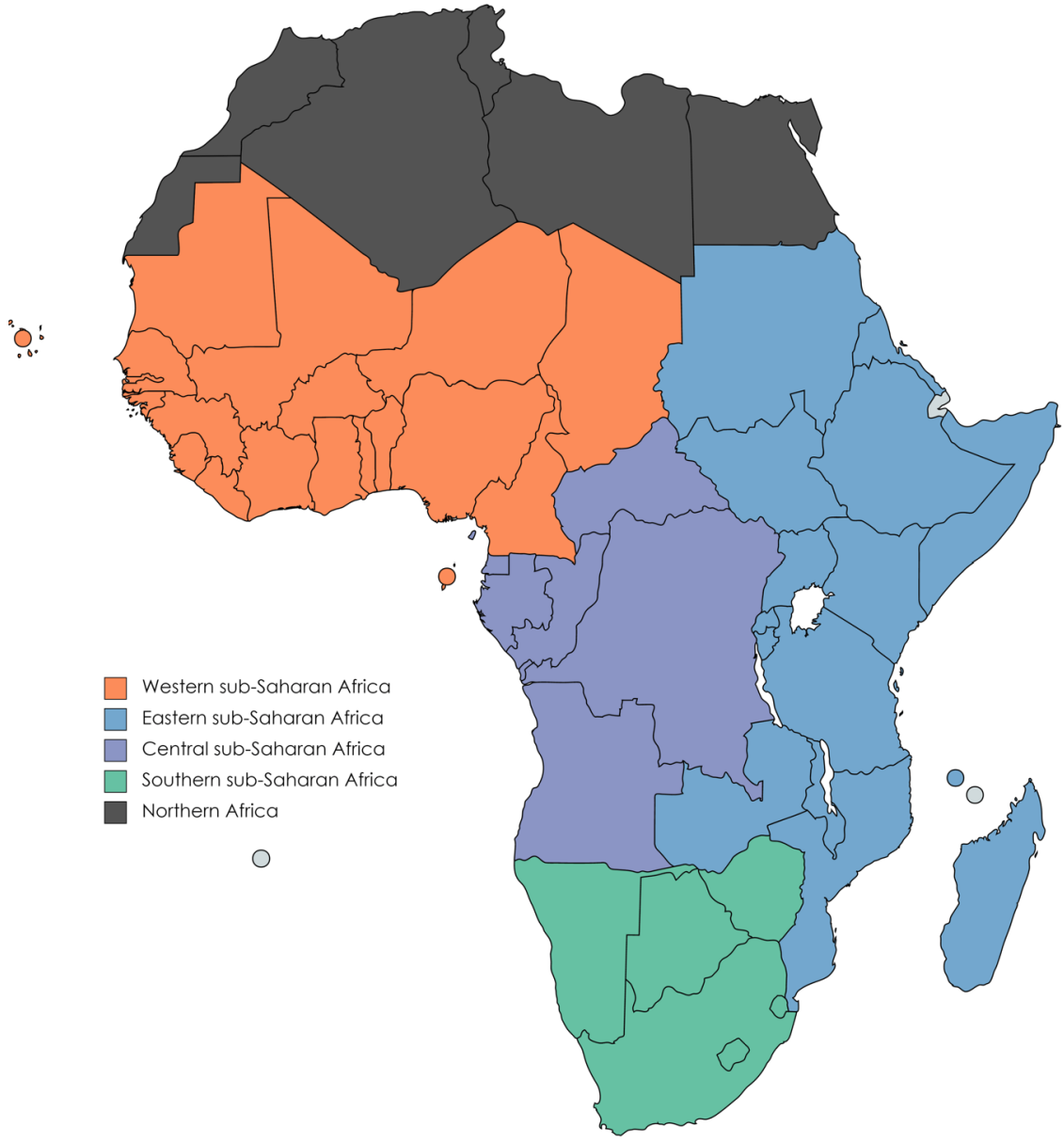

Figure 1.2. Super-regions of sub-Saharan Africa defined according to the Global Health Data Exchange region and super region classification system.
Grouping of countries used for reviewing the literature involving the study of *Shigella* in sub-Saharan Africa (https://ghdx.healthdata.org/countries).

Due to the uneven level of research between regions across sub-Saharan Africa, we can gain a greater insight into the epidemiology of *Shigella* in the region by looking at studies by sub-region (Figure 1.4).

1.6.1. Prevalence of *Shigella* in sub-Saharan Africa

The proportion of diarrhoeal cases caused by *Shigella* varies greatly both within a country and between countries (Table 1.1). Some of this variety may be due to the lack of distinction between EIEC and *Shigella*, increasing the proportion attributed.

In Western sub-Saharan Africa (Figure 1.4) the studies which specifically detected *Shigella* found them in 0-8% of MSD cases while the studies which did not distinguish between EIEC and *Shigella* generally identified EIEC/*Shigella* in a greater proportion of MSD patients (Table 1.1). The difference between proportion of cases where *Shigella* vs EIEC/*Shigella* are detected suggests that prevalence of EIEC is high in this region of sub-Saharan Africa.

In Eastern sub-Saharan Africa (Figure 1.4), however, there does not appear to be the same difference in the proportion of *Shigella* positive MSD cases or controls compared to the proportion of EIEC/*Shigella* positive (Table 1.1). The prevalence of EIEC is perhaps low in this region which would improve the accuracy of *Shigella* prevalence estimates based on EIEC/*Shigella* detection rates.

The variance in the proportion of *Shigella* detection is highly diverse in Eastern sub-Saharan Africa, from 0-3.3% in Tanzania to 20.6% in Somalia (Table 1.1). Differences in levels of *Shigella* detection may well be due to differences in prevalence between countries, though a meta-analysis of studies conducted in Ethiopia show that prevalence can vary greatly across regions within the same country, and this likely also plays a large role (Table 1.1) [86].

Table 1.1. Prevalence of Shigella across sub-Saharan Africa

Summary of published papers 2015 to 2022 identified using Scopus publications web search.

Region	Study population	Detection level	Shigella or Shigella/EIEC?	Study details	Reference
		Western sub-Saharan Africa			
Guinea-Bissau	Cases	63.3%	*Shigella/EIEC*	<5 years old, 2010-2012	[87]
Côte d'Ivoire	Cases	11-20%	*Shigella/EIEC*	All ages, 2012 and 2013	[88, 89]
The Gambia	Cases	9.3%	*Shigella/EIEC*	<5 years, 2008-2010	[14]
Ghana	Cases	0-1.2%	*Shigella*	All ages, 2010-2012	[90, 91]
Ghana	Cases	30.5%	*Shigella/EIEC*	<14 years, 2007-2008	[92]
Burkina Faso	Cases	8%	*Shigella*	<5 years, 2006-2008	[93]
Ghana	Asymptomatic controls	24.6%	*Shigella/EIEC*	<14 years, 2007-2008	[92]
The Gambia	Asymptomatic controls	0%	*Shigella/EIEC*	Adults, 2012	[94]
		Eastern sub-Saharan Africa			
Ethiopia	Cases	1.1-34.6%	Mixed	Meta-analysis: published 1999-2018	[86]
Somalia	Cases	20.6%	*Shigella*	<5 years, 2019	[95]
Mwanza, Tanzania	Cases	3.3%	*Shigella*	<5 years, 2015-2016	[96]
Tanzania	Cases	0%	*Shigella*	MAL-ED study, <5 years	[97]
Tanzania	Cases	14.5%	*Shigella/EIEC*	All ages, 2010-2015	[98]
Kenya	Cases	3.2-20%	*Shigella*	<5 years, 2008-2014	[14, 99-104]
Malawi	Cases	15%	*Shigella/EIEC*		[105, 106]
Madagascar	Cases	1.5%	*Shigella*	<5 years, 2011-2014	[107]
Mozambique	Cases	1.5-6.1%	*Shigella*	Rural, <5 years, 2007-2012	[14, 108, 109]
Khartoum, Sudan	Cases	~8%	*Shigella*	<5 years, 2013 and 2014	[110, 111]
Zambia	Cases	14.4%	*Shigella*	<5 years, 2016	[112]
Ethiopia	Asymptomatic controls	0.0-3.1%	Mixed	Meta-analysis: published 1999-2018	[86]
Juba, South Sudan	Healthy individuals	14%	*Shigella/EIEC*	Displaced persons camp, all ages	[113]
Mozambique	Asymptomatic controls	0.2-1%	*Shigella*	Rural, <5 years, 2007-2012	[14, 108, 109]
		Central sub-Saharan Africa			
The Democratic Republic of the Congo	Bloodstream infections	5.8%	*Shigella*		[114]
		Southern Sub-Saharan Africa			
Botswana	Cases	17-27%	*Shigella/EIEC*	Children, 2011-2014	[115-117]

There is also variance in the estimates of asymptomatic carriage in both Western and Eastern sub-Saharan Africa (Table 1.1). The level of asymptomatic carriage is likely to be influenced by the study population, for example asymptomatic carriage in healthy individuals in a displaced persons camp receiving vaccinations is likely to be higher, due to the conditions in such camps, compared to the levels found in a study of healthy food workers [86, 113]. No recent asymptomatic carriage data exists for Central or Southern sub-Saharan Africa, in fact very little recent data regarding the prevalence or contribution to diarrhoeal disease burden of *Shigella* exists for either region (Table 1.1).

1.6.2. Serodiversity of *Shigella* in sub-Saharan Africa

The recent studies of *Shigella* in sub-Saharan Africa, in which serogroup is identified, indicate that *S. flexneri* is the dominant serogroup across sub-Saharan Africa (Table 1.2). This fits with expectations based on the known global epidemiology and the known association between *S. sonnei* and industrialisation. Strains from all other serogroups have also been identified across sub-Saharan Africa (Table 1.2).

The *S. sonnei* serogroup has only a single serotype and is therefore likely to have a relatively high prevalence, despite the serogroup typically making up a minority of the identified *Shigella*, as the prevalence of *S. flexneri* is likely a composite of multiple serotypes. However, most studies do not test for a specific serotype, so the relative prevalence of specific serotypes is unknown.

Table 1.2. Serodiversity of Shigella across sub-Saharan Africa

Summary of published papers 2015 to 2022 identified using Scopus publications web search.

Region	Serogroup	Proportion of *Shigella*	Serotypes	Detection level	Study details	Reference
Western sub-Saharan Africa						
Niger	*S. flexneri*	56.5%	-	-	Rotavirus surveillance study	[118]
	S. dysenteriae	19.7%	-	-		
	S. boydii	17.2%	-	-		
	S. sonnei	6.6%	-	-		
Sierra Leone	*S. flexneri*	-	-	-	Aetiological agents of fever	[119]
	S. sonnei	-	-	-		
Eastern sub-Saharan Africa						
Somalia	*S. flexneri*	70%	-	-	Shigellosis study	[95]
	S. sonnei	-	-	-		
	S. dysenteriae	-	-	-		
Ethiopia	*S. flexneri*	52%	-	-	Shigellosis study	[120]
	S. sonnei	32%	-	-		
	S. boydii	16%	-	-		
Malawi	*S. flexneri*	100%	-	-	Shigellosis study	[106]
Kenya	*S. flexneri*	0-60%	-	-	MSD studies and shigellosis studies	[100-102, 104, 121]
	S. sonnei	12-46%	-	-		
	S. boydii	12-57%	-	-		
	S. dysenteriae	0-17%	-	-		
Mozambique	*S. flexneri*	70.1%	2a 6 1b Others	26.9-38% 13.4-18.9% 10.4-10.8% -	MSD study and shigellosis study	Serogroups: [122] Serotypes: [122, 123]
	S. sonnei	23.9%	-	-		
	S. boydii	3%	-	-		
Central sub-Saharan Africa						
Central African Republic	*S. flexneri*	55%	-	-	Shigellosis study	[124]
	S. sonnei	20%	-	-		
	S. dysenteriae	16%	-	-		
	S. boydii	8%	-	-		
Gabon	*S. flexneri*	~78%	-	-	Shigellosis study	[125]
	S. boydii	13.5%	-	-		
	S. sonnei	8.1%	-	-		
Southern Sub-Saharan Africa						
South Africa	*S. flexneri*	17%	-	-	Enteric bacteria study	[126]

1.6.3. Antimicrobial resistance of *Shigella* in sub-Saharan Africa

The literature suggests that the global trend of widespread MDR in *Shigella* is also true of sub-Saharan Africa (Table 1.3). While WGSA points to FQR strains emerging in Asia, resistant strains have been identified phenotypically in sub-Saharan Africa (Table 1.3) [78]. While it is not known if these FQR African strains are locally emerging FQR strains or imported FQR strains, it does appear that the prevalence of these strains remains low (Table 1.3) [86, 127].

Table 1.3. Antimicrobial resistance phenotypes of Shigella across sub-Saharan Africa

Summary of published papers 2015 to 2022 identified using Scopus publications web search. No data was found for western or southern sub-Saharan Africa.

Region	Antimicrobial resistance	Proportion of *Shigella*	Extra details	Reference
Eastern sub-Saharan Africa				
Ethiopia	Ampicillin, erythromycin, tetracycline, cotrimoxazole and chloramphenicol	>50% 83.3% each antimicrobial	Metanalysis	[86]
	Ciprofloxacin	8.9%		
	Ceftriaxone	9.3%		
Mozambique	Cotrimoxazole	92.5%		[123]
	Tetracycline	68.7%		
	Chloramphenicol	53.7%		
	Ampicillin	50.7%		
	MDR	S. flexneri: 47.7% Others: 7.4%		
	Cephalosporins, quinolones or aminoglycosides	0%		
Zambia	MDR	100%		[112]
	Ampicillin	*S. flexneri, S. boydii* and *S. dysenteriae*: 100%		
	Co-trimoxazole	*S. flexneri* and *S. boydii*: 100% *S. dysenteriae*: 75%		
	Chloramphenicol	*S. flexneri*: 83.8% *S. boydii*: 100% *S. dysenteriae*: 25%		
	Streptomycin	*S. flexneri*: 83.8%		
	Amoxicillin-clavulanic acid	*S. flexneri*: 16.5%		
	Tetracycline	*S. flexneri*: 16.5% *S. boydii*: 100% *S. dysenteriae*: 25%		
Central sub-Saharan Africa				
Central African Republic	Tetracycline, cotrimoxazole and sulfonamides	>90% 83.3% each antimicrobial		[47]
	Amoxicillin and ticarcillin	>65% 83.3% each antimicrobial (<20% S. sonnei)		
The Democratic Republic of the Congo	MDR	66.7%	Bloodstream isolates	[114]
	Cotrimoxazole, gentamycin, erythromycin and cefuroxime	100%		
	Ceftriaxone, ceftazidime, cefuroxime and ampicillin	83.3% of each antimicrobial		
	Amoxicillin	50%		
	Ciprofloxacin	33.3%		

Ampicillin resistance appears to be high across sub-Saharan Africa, while resistance to other antimicrobials is more varied between regions (Table 1.3) [127]. Resistance in *S. flexneri* appears to be greater than in *S. sonnei*, with *S. flexneri* strains being more likely to be MDR (Table 1.3) [123, 124]. Little work has been done to characterise the responsible AMR determinants, so genetic similarity/disparity in AMR genetic determinants across sub-Saharan Africa is unknown.

1.7. Thesis aims and objectives

This thesis aims to provide useful insights into African shigellosis, aiding the development of effective healthcare policy, both within sub-Sharan Africa and across the globe. This work also aims to begin addressing the under-utilisation of WGSA for studying African shigellosis, creating a more complete understanding of the global epidemiology of *Shigella,* and promoting further research into shigellosis in the region. These aims will be achieved through studying the shigellosis at different "levels"; local, national, and sub-continental, with a focus on antimicrobial resistance.

1.7.1. Thesis overview

The local level study examines *Shigella* isolates collected from children under five years old presenting with moderate-to-severe diarrhoea at the Queen Elizabeth hospital in Blantyre, Malawi. The study provides insights into *Shigella* diversity within a large urban hospital catchment area in Eastern Africa.

The national level studies focus on the two most prevalent serotypes in South Africa, *Shigella flexneri* 2a and *Shigella sonnei*, using isolates collected as part of routine surveillance of shigellosis in public health hospitals. This in-depth examination of the national epidemiology of these two *Shigella* serotypes identifies some of the factors influencing *Shigella* epidemiology, on both a national and international level and highlights the lifestyle and epidemiological differences between the two most prevalent *Shigella* serogroups.

The final study brings together the local and national study isolates with a previously characterised African *Shigella* dataset, collected as part of the Global Enteric Multicentre study (GEMS). This study provides insights into the epidemiology of *Shigella* across sub-Saharan Africa.

Chapter 2

Methods

Preface

Methods common across all the results chapters (Chapters 3, 4, 5, and 6) are detailed here. This chapter also outlines some AMR genotype-phenotype comparison work which informed AMR resistance phenotype prediction. The contributions of my various collaborators to the methods discussed are laid out in the table below, with study specific contributions from my collaborators defined within the respective study chapters.

Khuzwayo C. Jere	Assisted with the collection of isolates from Malawi [128].
Chikondi Peno	Assisted with the collection of isolates from Malawi [128].
End Chinyama	Assisted with the collection of isolates from Malawi [128].
Jonathan Mandolo	Assisted with the collection of isolates from Malawi [128].
Naor Bar-Zeev	Assisted with the collection of isolates from Malawi [128].
Amy K. Cain	Assisted with the collection of isolates from Malawi [128].
Nigel Cunliffe	Sequenced the Malawian isolates [128].
Karen Keddy	Collected the South African isolates.
Anthony Smith	Provided metadata for South African isolates
Juno Thomas	Provided metadata for South African isolates
Ross Low (Earlham Institute)	South African isolates whole genome sequencing GEMS study isolates whole genome sequencing [129]
Neil Hall (Earlham Institute)	South African isolates whole genome sequencing GEMS study isolates whole genome sequencing [129]
Centre for Genomics Research	South African and GEMS study resequencing
Caisey J. Pulford	Provided an optimised whole genome sequence read quality trimming pipeline
Rebecca J. Bengtsson	Provided accession numbers for the phylogenetic reference isolates, having created known global population structure-representative, serogroup isolate lists, and smaller serogroup representative isolate lists. Also provided a working core-SNP alignment generation pipeline and bioinformatics support.

2.1. Samples

The studies in this thesis use sample collections of different sizes, sampled from varying population sizes.

2.1.1. Malawian isolates

In Chapter 3, a study of the genomic epidemiology of *Shigella* within a single hospital catchment area, all the biochemically identified shigellae collected during a rotavirus vaccine efficacy study in Blantyre, Malawi, were included in the sample set [105, 128]. Samples (n = 10) were collected from children under the age of five years attending the Queen Elizabeth Central Hospital with acute gastroenteritis between 2012 to 2015 [105, 128].

2.1.2. South African isolates

For Chapters 4 and 5, serotype specific studies examining the genomic epidemiology of *Shigella* on a national level, all samples (n = 561) were collected as part of routine shigellosis surveillance in South Africa from 2011 to 2015. Surveillance was carried out by The Group for Enteric, Respiratory and Meningeal Diseases Surveillance in South Africa (GERMS-SA) and involved a network of public hospitals and laboratories across all nine provinces. Isolates were collected across all age groups but only biochemically identified *Shigella flexneri* 2a or *Shigella sonnei* were selected. The isolates were selected as a single set and subsequently separated by serotype, with *S. flexneri* 2a studied in Chapter 4 and *S. sonnei* in Chapter 5.

2.1.3. Global enteric multicentre study isolates

Chapter 6 brings together multiple *Shigella* sample sets from across sub-Saharan Africa. All *Shigella* isolates from Chapters 3, 4 and 5 are examined in the context of a previously studied African *Shigella* sample set, collected during the Global Enteric Multicentre Study (GEMS) [129]. The GEMS study isolates were collected from children under the age of five attending a public hospital with acute gastroenteritis. The isolates were collected across thirty healthcare centres in four African countries (namely, Mali, The Gambia, Mozambique, and Kenya; 9, 5, 5 and 11 healthcare centres respectively)

from 2007 to 2011 [129]. All biochemically or *in silico* identified *S. flexneri* and *S. sonnei* collected at African sites were included in the Chapter 6 multinational study. Details on *in silico* typing are described in Section 2.5 and specifics on how this was used to define the study set are described in Chapter 6.

2.1.4. Population structure reference isolates

Phylogenetic trees were created for each results chapter, methods detailed in Section 2.5.1, and a collection of whole genome sequenced reference isolates were included to provide context and population structure for the study isolates (Tables 2.1, 2.2, and 2.3).

For Chapter 3, isolates were selected from across the known *Escherichia coli/Shigella* clade, including all *Shigella* serogroups (Table 2.1). While at least one Isolate was selected from all currently identified *E. coli/Shigella* Phylogroups, initially, some isolates were excluded for failing to meet read quality thresholds and so not all phylogroups were represented in the final tree. New isolates were not selected as replacements for the excluded isolates as the identification of Phylogroup was not the aim of creating the tree.

Two separate reference sample sets were needed for Chapters 4, 5 and 6, one for *S. flexneri* (Table 2.2) and one for *S. sonnei* (Table 2.3). For both sample sets, isolates were selected from across the known global Phylogeny, with care taken to include isolates from all Phylogroups/Lineages and serotypes (carried out by Dr Bengtsson, collaborator contributions above).

Table 2.1. Maximum likelihood phylogenetics, Escherichia coli/Shigella clade population structure reference genomes.

Accession number	Species	Phylogroup	Chapter	Reference
SRR3234362	*S. boydii*	2	3	[36]
SRR3234365	*S. boydii*	1	3	[36]
SRR3237801	*S. boydii*	2	3	[36]
SRR3237803	*S. boydii*	1	3	[36]
SRR3237805	*S. boydii*	1	3	[36]
SRR3237806	*S. boydii*	2	3	[36]
SRR3237808	*S. boydii*	1	3	[36]
SRR3237809	*S. boydii*	2	3	[36]
SRR3237810	*S. boydii*	2	3	[36]
SRR3237816	*S. boydii*	3	3	[36]
SRR3237817	*S. boydii*	3	3	[36]
SRR3237837	*S. boydii*	2	3	[36]
SRR3237947	*S. boydii*	3	3	[36]
SRR3237948	*S. boydii*	2	3	[36]
SRR3237949	*S. boydii*	2	3	[36]
SRR3237950	*S. boydii*	3	3	[36]
SRR3237951	*S. boydii*	3	3	[36]
SRR3237953	*S. boydii*	2	3	[36]
SRR3237955	*S. boydii*	3	3	[36]
SRR3237957	*S. boydii*	3	3	[36]
SRR3237959	*S. boydii*	3	3	[36]
SRR3237960	*S. boydii*	3	3	[36]
SRR3237961	*S. boydii*	1	3	[36]
SRR3237962	*S. boydii*	1	3	[36]
SRR3237963	*S. boydii*	2	3	[36]
SRR3237964	*S. boydii*	3	3	[36]
SRR3237965	*S. boydii*	3	3	[36]
ERR1013817	*S. dysenteriae*	II	3	[34]
ERR1013857	*S. dysenteriae*	II	3	[34]
ERR1014002	*S. dysenteriae*	IV	3	[34]
ERR1014032	*S. dysenteriae*	II	3	[34]
ERR1014042	*S. dysenteriae*	IV	3	[34]
ERR1014154	*S. dysenteriae*	III	3	[34]
ERR1014501	*S. dysenteriae*	III	3	[34]
ERR1014536	*S. dysenteriae*	I	3	[34]
ERR1014544	*S. dysenteriae*	III	3	[34]
ERR1014555	*S. dysenteriae*	IV	3	[34]
SRR072242	*E. coli*	A	3	
SRR10997234	*E. coli*	A	3	
SRR1509643	*E. coli*	E	3	[130]
SRR2061820	*E. coli*	B2	3	[131]
SRR2169510	*E. coli*	B2	3	
SRR2169556	*E. coli*	B1	3	
SRR5825845	*E. coli*	B1	3	
ERR042803	*S. flexneri*	3	3	[72]
ERR042850	*S. flexneri*	3	3	[72]

ERR048281	*S. flexneri*	2	3	[33]
ERR048288	*S. flexneri*	6	3	[33]
ERR048302	*S. flexneri*	3	3	[72]
ERR048305	*S. flexneri*	1	3	[33]
ERR048317	*S. flexneri*	7	3	[33]
ERR048339	*S. flexneri*	3	3	[72]
ERR126987	*S. flexneri*	3	3	[72]
ERR126993	*S. flexneri*	2	3	[33]
ERR127032	*S. flexneri*		3	[132]
ERR127033	*S. flexneri*		3	[132]
ERR127034	*S. flexneri*		3	[132]
ERR127035	*S. flexneri*		3	[132]
ERR127036	*S. flexneri*		3	[132]
ERR127037	*S. flexneri*		3	[132]
ERR127038	*S. flexneri*		3	[132]
ERR127039	*S. flexneri*		3	[132]
ERR127040	*S. flexneri*		3	[132]
ERR127041	*S. flexneri*		3	[132]
ERR127043	*S. flexneri*		3	[132]
ERR127044	*S. flexneri*		3	[132]
ERR127045	*S. flexneri*		3	[132]
ERR127046	*S. flexneri*		3	[132]
ERR127047	*S. flexneri*		3	[132]
ERR127048	*S. flexneri*		3	[132]
ERR1363976	*S. flexneri*	3	3	[72]
ERR1364007	*S. flexneri*	3	3	[72]
ERR1364014	*S. flexneri*	3	3	[72]
ERR1364050	*S. flexneri*	3	3	[33]
ERR1364087	*S. flexneri*	3	3	[33]
ERR1364097	*S. flexneri*	3	3	[33]
ERR1364106	*S. flexneri*	3	3	[33]
ERR1364137	*S. flexneri*	3	3	[33]
ERR200376	*S. flexneri*	3	3	[33]
ERR217085	*S. flexneri*	1	3	[33]
ERR449043	*S. flexneri*		3	[133]
ERR449077	*S. flexneri*		3	[133]
ERR559526	*S. flexneri*		3	[134]
ERR832464	*S. flexneri*	5	3	[33]
ERR832481	*S. flexneri*	3	3	[33]
SRR7886341	*S. flexneri*		3	

Table 2.2. Maximum likelihood phylogenetics, Shigella flexneri global population structure reference genomes.

Accession number	Species	Phylogroup	Chapter	Reference
ERS025897	*S. flexneri*	3	4, 5, 6	
ERS025898	*S. flexneri*	4	4, 5, 6	
ERS025900	*S. flexneri*	3	4, 5, 6	
ERS025902	*S. flexneri*	3	4, 5, 6	
ERS025903	*S. flexneri*	3	4, 5, 6	
ERS025904	*S. flexneri*	3	4, 5, 6	
ERS025905	*S. flexneri*	3	4, 5, 6	
ERS025906	*S. flexneri*	3	4, 5, 6	
ERS025907	*S. flexneri*	3	4, 5, 6	
ERS025908	*S. flexneri*	1	4, 5, 6	
ERS025910	*S. flexneri*	1	4, 5, 6	
ERS025911	*S. flexneri*	1	4, 5, 6	
ERS025912	*S. flexneri*	2	4, 5, 6	
ERS025913	*S. flexneri*	1	4, 5, 6	
ERS025914	*S. flexneri*	1	4, 5, 6	
ERS025915	*S. flexneri*	3	4, 5, 6	
ERS025916	*S. flexneri*	3	4, 5, 6	
ERS025917	*S. flexneri*	3	4, 5, 6	
ERS025919	*S. flexneri*	1	4, 5, 6	
ERS025920	*S. flexneri*	2	4, 5, 6	
ERS025922	*S. flexneri*	3	4, 5, 6	
ERS025923	*S. flexneri*	2	4, 5, 6	
ERS025925	*S. flexneri*	3	4, 5, 6	
ERS025926	*S. flexneri*	3	4, 5, 6	
ERS025928	*S. flexneri*	7	4, 5, 6	
ERS025929	*S. flexneri*	1	4, 5, 6	
ERS025930	*S. flexneri*	2	4, 5, 6	
ERS025931	*S. flexneri*	3	4, 5, 6	
ERS025932	*S. flexneri*	1	4, 5, 6	
ERS025933	*S. flexneri*	6	4, 5, 6	
ERS025934	*S. flexneri*	6	4, 5, 6	
ERS025936	*S. flexneri*	1	4, 5, 6	
ERS025937	*S. flexneri*	2	4, 5, 6	
ERS025938	*S. flexneri*	2	4, 5, 6	
ERS025939	*S. flexneri*	7	4, 5, 6	
ERS025940	*S. flexneri*	3	4, 5, 6	
ERS025941	*S. flexneri*	3	4, 5, 6	
ERS025942	*S. flexneri*	3	4, 5, 6	
ERS025943	*S. flexneri*	3	4, 5, 6	
ERS025944	*S. flexneri*	4	4, 5, 6	
ERS025946	*S. flexneri*	3	4, 5, 6	
ERS025947	*S. flexneri*	1	4, 5, 6	
ERS025949	*S. flexneri*	2	4, 5, 6	
ERS025950	*S. flexneri*	1	4, 5, 6	
ERS025951	*S. flexneri*	3	4, 5, 6	
ERS025952	*S. flexneri*	1	4, 5, 6	
ERS025953	*S. flexneri*	1	4, 5, 6	

ERS025954	*S. flexneri*	1	4, 5, 6
ERS025955	*S. flexneri*	3	4, 5, 6
ERS025956	*S. flexneri*	2	4, 5, 6
ERS025959	*S. flexneri*	7	4, 5, 6
ERS025961	*S. flexneri*	3	4, 5, 6
ERS025962	*S. flexneri*	3	4, 5, 6
ERS025964	*S. flexneri*	3	4, 5, 6
ERS033311	*S. flexneri*	4	4, 5, 6
ERS033313	*S. flexneri*	1	4, 5, 6
ERS033317	*S. flexneri*	5	4, 5, 6
ERS033318	*S. flexneri*	5	4, 5, 6
ERS033324	*S. flexneri*	5	4, 5, 6
ERS033325	*S. flexneri*	5	4, 5, 6
ERS033327	*S. flexneri*	3	4, 5, 6
ERS033331	*S. flexneri*	3	4, 5, 6
ERS033333	*S. flexneri*	2	4, 5, 6
ERS033335	*S. flexneri*	3	4, 5, 6
ERS033336	*S. flexneri*	3	4, 5, 6
ERS033337	*S. flexneri*	4	4, 5, 6
ERS033338	*S. flexneri*	1	4, 5, 6
ERS033340	*S. flexneri*	6	4, 5, 6
ERS033342	*S. flexneri*	1	4, 5, 6
ERS033343	*S. flexneri*	1	4, 5, 6
ERS033344	*S. flexneri*	3	4, 5, 6
ERS033345	*S. flexneri*	2	4, 5, 6
ERS033346	*S. flexneri*	2	4, 5, 6
ERS033347	*S. flexneri*	3	4, 5, 6
ERS033348	*S. flexneri*	2	4, 5, 6
ERS033351	*S. flexneri*	1	4, 5, 6
ERS033352	*S. flexneri*	4	4, 5, 6
ERS033353	*S. flexneri*	3	4, 5, 6
ERS033354	*S. flexneri*	3	4, 5, 6
ERS033355	*S. flexneri*	3	4, 5, 6
ERS033356	*S. flexneri*	3	4, 5, 6
ERS033357	*S. flexneri*	1	4, 5, 6
ERS033358	*S. flexneri*	3	4, 5, 6
ERS033360	*S. flexneri*	7	4, 5, 6
ERS033361	*S. flexneri*	7	4, 5, 6
ERS033362	*S. flexneri*	3	4, 5, 6
ERS033363	*S. flexneri*	7	4, 5, 6
ERS033364	*S. flexneri*	7	4, 5, 6
ERS033365	*S. flexneri*	3	4, 5, 6
ERS033366	*S. flexneri*	1	4, 5, 6
ERS033369	*S. flexneri*	7	4, 5, 6
ERS033370	*S. flexneri*	3	4, 5, 6
ERS033371	*S. flexneri*	3	4, 5, 6
ERS033372	*S. flexneri*	3	4, 5, 6
ERS033373	*S. flexneri*	3	4, 5, 6
ERS033374	*S. flexneri*	3	4, 5, 6

ERS033375	*S. flexneri*	3	4, 5, 6
ERS033376	*S. flexneri*	3	4, 5, 6
ERS033379	*S. flexneri*	3	4, 5, 6
ERS033380	*S. flexneri*	7	4, 5, 6
ERS033381	*S. flexneri*	7	4, 5, 6
ERS033382	*S. flexneri*	3	4, 5, 6
ERS033383	*S. flexneri*	3	4, 5, 6
ERS033384	*S. flexneri*	7	4, 5, 6
ERS033385	*S. flexneri*	7	4, 5, 6
ERS033386	*S. flexneri*	7	4, 5, 6
ERS033387	*S. flexneri*	5	4, 5, 6
ERS033388	*S. flexneri*	3	4, 5, 6
ERS033389	*S. flexneri*	3	4, 5, 6
ERS033390	*S. flexneri*	3	4, 5, 6
ERS033391	*S. flexneri*	3	4, 5, 6
ERS087985	*S. flexneri*	3	4, 5, 6
ERS087986	*S. flexneri*	3	4, 5, 6
ERS087988	*S. flexneri*	3	4, 5, 6
ERS087989	*S. flexneri*	3	4, 5, 6
ERS087991	*S. flexneri*	2	4, 5, 6
ERS087993	*S. flexneri*	2	4, 5, 6
ERS087994	*S. flexneri*	2	4, 5, 6
ERS087995	*S. flexneri*	2	4, 5, 6
ERS088014	*S. flexneri*	3	4, 5, 6
ERS088015	*S. flexneri*	3	4, 5, 6
ERS088016	*S. flexneri*	3	4, 5, 6
ERS088017	*S. flexneri*	3	4, 5, 6
ERS088018	*S. flexneri*	3	4, 5, 6
ERS088019	*S. flexneri*	2	4, 5, 6
ERS088020	*S. flexneri*	2	4, 5, 6
ERS088021	*S. flexneri*	2	4, 5, 6
ERS088022	*S. flexneri*	2	4, 5, 6
ERS088023	*S. flexneri*	2	4, 5, 6
ERS088040	*S. flexneri*	3	4, 5, 6
ERS088041	*S. flexneri*	3	4, 5, 6
ERS088042	*S. flexneri*	3	4, 5, 6
ERS088043	*S. flexneri*	3	4, 5, 6
ERS088044	*S. flexneri*	3	4, 5, 6
ERS088045	*S. flexneri*	3	4, 5, 6
ERS088046	*S. flexneri*	2	4, 5, 6
ERS088047	*S. flexneri*	2	4, 5, 6
ERS088048	*S. flexneri*	2	4, 5, 6
ERS088049	*S. flexneri*	2	4, 5, 6
ERS088050	*S. flexneri*	2	4, 5, 6
ERS088060	*S. flexneri*	1	4, 5, 6
ERS088061	*S. flexneri*	1	4, 5, 6
ERS088062	*S. flexneri*	1	4, 5, 6
ERS088063	*S. flexneri*	3	4, 5, 6
ERS088064	*S. flexneri*	1	4, 5, 6

ERS088065	*S. flexneri*	4	4, 5, 6
ERS088066	*S. flexneri*	4	4, 5, 6
ERS088067	*S. flexneri*	4	4, 5, 6
ERS088068	*S. flexneri*	1	4, 5, 6
ERS088069	*S. flexneri*	1	4, 5, 6
ERS088070	*S. flexneri*	5	4, 5, 6
ERS088071	*S. flexneri*	5	4, 5, 6
ERS088072	*S. flexneri*	5	4, 5, 6
ERS088074	*S. flexneri*	4	4, 5, 6
ERS088075	*S. flexneri*	3	4, 5, 6
ERS093670	*S. flexneri*	3	4, 5, 6
ERS093671	*S. flexneri*	1	4, 5, 6
ERS093673	*S. flexneri*	3	4, 5, 6
ERS093674	*S. flexneri*	3	4, 5, 6
ERS093675	*S. flexneri*	3	4, 5, 6
ERS093676	*S. flexneri*	5	4, 5, 6
ERS093677	*S. flexneri*	1	4, 5, 6
ERS093679	*S. flexneri*	3	4, 5, 6
ERS093680	*S. flexneri*	3	4, 5, 6
ERS093682	*S. flexneri*	2	4, 5, 6
ERS093683	*S. flexneri*	1	4, 5, 6
ERS093684	*S. flexneri*	3	4, 5, 6
ERS093685	*S. flexneri*	3	4, 5, 6
ERS093686	*S. flexneri*	2	4, 5, 6
ERS093687	*S. flexneri*	1	4, 5, 6
ERS093688	*S. flexneri*	1	4, 5, 6
ERS093689	*S. flexneri*	1	4, 5, 6
ERS093690	*S. flexneri*	4	4, 5, 6
ERS093691	*S. flexneri*	3	4, 5, 6
ERS093692	*S. flexneri*	1	4, 5, 6
ERS093693	*S. flexneri*	1	4, 5, 6
ERS093694	*S. flexneri*	4	4, 5, 6
ERS093696	*S. flexneri*	1	4, 5, 6
ERS093698	*S. flexneri*	3	4, 5, 6
ERS093699	*S. flexneri*	1	4, 5, 6
ERS093700	*S. flexneri*	1	4, 5, 6
ERS093702	*S. flexneri*	3	4, 5, 6
ERS093703	*S. flexneri*	1	4, 5, 6
ERS093704	*S. flexneri*	3	4, 5, 6
ERS093705	*S. flexneri*	4	4, 5, 6
ERS093706	*S. flexneri*	3	4, 5, 6
ERS093707	*S. flexneri*	4	4, 5, 6
ERS093708	*S. flexneri*	1	4, 5, 6
ERS093709	*S. flexneri*	4	4, 5, 6
ERS093712	*S. flexneri*	3	4, 5, 6
ERS157637	*S. flexneri*	3	4, 5, 6
ERS157638	*S. flexneri*	3	4, 5, 6
ERS157639	*S. flexneri*	3	4, 5, 6
ERS157640	*S. flexneri*	3	4, 5, 6

ERS157641	*S. flexneri*	3	4, 5, 6
ERS157642	*S. flexneri*	3	4, 5, 6
ERS157643	*S. flexneri*	3	4, 5, 6
ERS157644	*S. flexneri*	3	4, 5, 6
ERS157645	*S. flexneri*	2	4, 5, 6
ERS157646	*S. flexneri*	3	4, 5, 6
ERS157647	*S. flexneri*	3	4, 5, 6
ERS157648	*S. flexneri*	3	4, 5, 6
ERS157649	*S. flexneri*	3	4, 5, 6
ERS157650	*S. flexneri*	3	4, 5, 6
ERS157651	*S. flexneri*	3	4, 5, 6
ERS157652	*S. flexneri*	3	4, 5, 6
ERS157653	*S. flexneri*	3	4, 5, 6
ERS157654	*S. flexneri*	3	4, 5, 6
ERS157655	*S. flexneri*	3	4, 5, 6
ERS157656	*S. flexneri*	3	4, 5, 6
ERS157657	*S. flexneri*	3	4, 5, 6
ERS157658	*S. flexneri*	3	4, 5, 6
ERS157659	*S. flexneri*	3	4, 5, 6
ERS157660	*S. flexneri*	3	4, 5, 6
ERS157661	*S. flexneri*	3	4, 5, 6
ERS157662	*S. flexneri*	3	4, 5, 6
ERS157663	*S. flexneri*	3	4, 5, 6
ERS157664	*S. flexneri*	3	4, 5, 6
ERS157665	*S. flexneri*	3	4, 5, 6
ERS157666	*S. flexneri*	3	4, 5, 6
ERS157667	*S. flexneri*	3	4, 5, 6
ERS157668	*S. flexneri*	3	4, 5, 6
ERS157669	*S. flexneri*	3	4, 5, 6
ERS157670	*S. flexneri*	2	4, 5, 6
ERS157671	*S. flexneri*	2	4, 5, 6
ERS157672	*S. flexneri*	2	4, 5, 6
ERS157673	*S. flexneri*	2	4, 5, 6
ERS157674	*S. flexneri*	2	4, 5, 6
ERS157675	*S. flexneri*	2	4, 5, 6
ERS157677	*S. flexneri*	2	4, 5, 6
ERS157678	*S. flexneri*	2	4, 5, 6
ERS157679	*S. flexneri*	2	4, 5, 6
ERS157680	*S. flexneri*	2	4, 5, 6
ERS157681	*S. flexneri*	2	4, 5, 6
ERS157682	*S. flexneri*	2	4, 5, 6
ERS157683	*S. flexneri*	2	4, 5, 6
ERS157684	*S. flexneri*	2	4, 5, 6
ERS157685	*S. flexneri*	2	4, 5, 6
ERS157686	*S. flexneri*	2	4, 5, 6
ERS157687	*S. flexneri*	2	4, 5, 6
ERS157688	*S. flexneri*	2	4, 5, 6
ERS157689	*S. flexneri*	2	4, 5, 6
ERS157690	*S. flexneri*	5	4, 5, 6

ERS157691	*S. flexneri*	6	4, 5, 6
ERS157692	*S. flexneri*	7	4, 5, 6
ERS157750	*S. flexneri*	6	4, 5, 6
ERS157751	*S. flexneri*	6	4, 5, 6
ERS157752	*S. flexneri*	6	4, 5, 6
ERS157753	*S. flexneri*	6	4, 5, 6
ERS157754	*S. flexneri*	6	4, 5, 6
ERS157755	*S. flexneri*	6	4, 5, 6
ERS157756	*S. flexneri*	6	4, 5, 6
ERS157757	*S. flexneri*	6	4, 5, 6
ERS157758	*S. flexneri*	2	4, 5, 6
ERS157759	*S. flexneri*	3	4, 5, 6
ERS157760	*S. flexneri*	6	4, 5, 6
ERS157761	*S. flexneri*	6	4, 5, 6
ERS157762	*S. flexneri*	6	4, 5, 6
ERS157763	*S. flexneri*	6	4, 5, 6
ERS157764	*S. flexneri*	6	4, 5, 6
ERS157765	*S. flexneri*	6	4, 5, 6
ERS157766	*S. flexneri*	6	4, 5, 6
ERS157767	*S. flexneri*	1	4, 5, 6
ERS157768	*S. flexneri*	6	4, 5, 6
ERS157769	*S. flexneri*	6	4, 5, 6
ERS157770	*S. flexneri*	6	4, 5, 6
ERS157771	*S. flexneri*	6	4, 5, 6
ERS157772	*S. flexneri*	6	4, 5, 6
ERS157773	*S. flexneri*	1	4, 5, 6
ERS157774	*S. flexneri*	6	4, 5, 6
ERS157785	*S. flexneri*	1	4, 5, 6
ERS157786	*S. flexneri*	1	4, 5, 6
ERS157787	*S. flexneri*	1	4, 5, 6
ERS157788	*S. flexneri*	3	4, 5, 6
ERS157813	*S. flexneri*	1	4, 5, 6
ERS157815	*S. flexneri*	1	4, 5, 6
ERS157816	*S. flexneri*	1	4, 5, 6
ERS157817	*S. flexneri*	1	4, 5, 6
ERS157818	*S. flexneri*	1	4, 5, 6
ERS157819	*S. flexneri*	1	4, 5, 6
ERS157820	*S. flexneri*	1	4, 5, 6
ERS157821	*S. flexneri*	1	4, 5, 6
ERS157822	*S. flexneri*	1	4, 5, 6
ERS157823	*S. flexneri*	1	4, 5, 6
ERS157824	*S. flexneri*	1	4, 5, 6
ERS157825	*S. flexneri*	1	4, 5, 6

Table 2.3. Maximum likelihood phylogenetics, Shigella sonnei global population structure reference genomes.

Accession number	Species	Phylogroup	Chapter	Reference
5008_7#11	*S. sonnei*	I	3, 4, 5, 6	[35]
5008_7#5	*S. sonnei*	III	3, 4, 5, 6	[35]
5008_7#6	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_1#5	*S. sonnei*	II	3, 4, 5, 6	[35]
5236_1#8	*S. sonnei*	II	3, 4, 5, 6	[35]
5236_2#12	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_2#3	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_2#4	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_2#5	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_5#10	*S. sonnei*		3, 4, 5, 6	[35]
5236_5#12	*S. sonnei*	II	3, 4, 5, 6	[35]
5236_5#3	*S. sonnei*	II	3, 4, 5, 6	[35]
5236_6#10	*S. sonnei*	II	3, 4, 5, 6	[35]
5236_6#2	*S. sonnei*	IV	3, 4, 5, 6	[35]
5236_7#10	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_7#11	*S. sonnei*		3, 4, 5, 6	[35]
5236_7#2	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_7#4	*S. sonnei*	II	3, 4, 5, 6	[35]
5236_7#7	*S. sonnei*	I	3, 4, 5, 6	[35]
5236_8#3	*S. sonnei*	II	3, 4, 5, 6	[35]
5236_8#6	*S. sonnei*	I	3, 4, 5, 6	[35]
5236_8#8	*S. sonnei*	III	3, 4, 5, 6	[35]
5236_8#9	*S. sonnei*	III	3, 4, 5, 6	[35]
5417_1#11	*S. sonnei*	II	3, 4, 5, 6	[35]
5417_1#2	*S. sonnei*	II	3, 4, 5, 6	[35]
5417_1#4	*S. sonnei*	III	3, 4, 5, 6	[35]
5417_1#6	*S. sonnei*	III	3, 4, 5, 6	[35]
5417_2#2	*S. sonnei*	III	3, 4, 5, 6	[35]
5417_2#9	*S. sonnei*	III	3, 4, 5, 6	[35]
5417_3#3	*S. sonnei*	II	3, 4, 5, 6	[35]
5417_3#8	*S. sonnei*	III	3, 4, 5, 6	[35]
8290_4#28	*S. sonnei*		3, 4, 5, 6	
8403_8#16	*S. sonnei*	III	3, 4, 5, 6	
8403_8#89	*S. sonnei*	V	3, 4, 5, 6	
8403_8#95	*S. sonnei*	V	3, 4, 5, 6	
8489_1#60	*S. sonnei*	III	3, 4, 5, 6	
9789_6#32	*S. sonnei*	III	3, 4, 5, 6	
9803_4#17	*S. sonnei*	III	3, 4, 5, 6	
9803_4#91	*S. sonnei*	III	3, 4, 5, 6	
9870_7#10	*S. sonnei*		3, 4, 5, 6	

2.2. Whole genome sequencing

All study isolates were subjected to whole genome sequencing (WGS). For Malawian isolates, Chapter 3, WGS was done at The Wellcome Trust Sanger Institute according to in-house protocols [135]. All South African (Chapters 4 and 5) and GEMS (Chapter 6) isolates were sequenced using Illumina HiSeq 4000 sequencing equipment and the DNA library was prepared using the Illumina Nextera XT DNA Library Prep Kit (Illumina, FC-131-1096) [136].

The sequence quality of some of the South African and GEMS isolates was poor and were re-sequenced at the Centre for Genomic Research (CGR, University of Liverpool) using the Illumina NovaSeq 6000 platform; the DNA library was constructed using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina [37]. Details on sequence read quality assessment are laid out in Section 2.4.1, while specifics on the number of re-sequenced isolates are in the relevant chapter.

The phylogenetic reference isolates were sequenced in previous studies using a range of technologies (Tables 2.1, 2.2 and 2.3).

2.3. Quality control

2.3.1. Sequence read quality

All raw sequence reads were quality trimmed with Trimmomatic (v0.38) and SeqTK (v1.3) (https://github.com/lh3/seqtk) [55]. In addition to quality trimming, a further 8 or 33 bases were removed from the ends of all South African study isolate reads and Malawian study isolate reads, respectively, due to poor quality.

Quality was assessed before and after trimming with fastQC (v0.11.8) and MultiQC (v1.5) [137, 138]. Read quality was assessed on 1) mean quality score, 2) per sequence quality score, 3) per base sequence content, 4) per sequence GC content, 5) per base N content, 6) overrepresented sequences, and 7) adapter content.

To be included in further analyses, isolate sequence reads had to meet these quality criteria: 1) per base mean Phred score ≥28, 2) per sequence Phred score ≥28, 3) even per base sequence content, 4) even per sequence GC content curve and peak GC content between 48-54%, 5) per base N content <5, 6) <5% overrepresented sequences, and 7) <5% sequences with adaptor sequences. Specifics on how many and why isolates were excluded for poor sequence quality are detailed in the relevant chapter.

2.3.2. Sequence read mapping quality

Read mapping quality was assessed using Qualimap (v2.2.2-dev) [139]. Any isolate with a mean read depth, across the whole genome, of <20 (South African isolates) or <10 (GEMS and phylogeny reference isolates) was excluded from further analysis. Mapping quality was performed by snippy (v4.3.6; https://github.com/tseemann/snippy) using default settings for the Malawian isolates.

2.3.3. Genome assembly quality

Genome assembly quality was assessed using Quast (v8.13) against the same reference genome to which sequence reads were mapped (Table 2.4) [140, 141]. Genome assembly methods are detailed in Section 2.6.

2.4. *In silico* strain typing

In silico confirmation as *Shigella* or enteroinvasive *Escherichia coli* (EIEC), and prediction of serotype was done using the *Shigella* typing software ShigaTyper (v1.0.6) (https://github.com/CFSAN-Biostatistics/shigatyper). Maximum likelihood phylogenetics was used to confirm serogroup *in silico*.

A phylogeny-based typing scheme having been devised for *S. sonnei*, I also used the sonneiTyping sonnei_genotype.py (v1) (https://github.com/katholt/sonneityping) to type my *S.* sonnei isolates according to this scheme [39]. The predicted Lineages and Clades were confirmed with maximum likelihood phylogenetics.

Both typing software, SonneiTyping and ShigeTyper, were run using isolate sequence reads, I used raw sequence reads with ShigaTyper but trimmed mapped reads with SonneiTyping.

2.4.1. Maximum likelihood phylogenetics

All maximum likelihood phylogenetic trees were generated from core SNP alignments using RAxML-NG (v0.6.6; GTR+G substitution model, 1000 bootstrap validation and mid-point rooted) [61]. Maximum likelihood phylogenies, in most cases, included a selection of reference isolates (Tables 2.1, 2.2 and 2.3) selected from across the known global phylogeny to provide context and tree structure for the study isolates (Section 2.2.4).

Core-SNP alignments were generated from quality trimmed sequenced reads by two separate methods. In the Malawian hospital study (Chapter 3) I used a more automated, single software method, while for the other chapters I used a more flexible multi-software pipeline approach.

The single software method used snippy (v4.3.6; https://github.com/tseemann/snippy) to map sequence reads to an *S. flexneri* complete reference genome (Table 2.2), filter out unmapped reads, call variant sites, and define a consensus sequence for each isolate.

The same process was achieved with the multi-software pipeline for the other chapters. Sequence read mapping, performed with bwa mem (v0.7.17), was to either the same *S. flexneri* complete

reference genome used in the snippy pipeline or a complete *S. sonnei* reference genome (Table 2.4). Samtools (v1.9) view (-F 3844 and -q 60) was used to remove all reads which were 1) unmapped, 2) not primarily aligned, 3) of insufficient quality, 4) mapped with insufficient quality (<60), 5) a PCR or optical duplicate, or 6) supplementarily aligned. Read alignments were clipped against the mapping reference genome with samclip (v2.27.1) and then sorted and indexed with samtools. Duplicate reads were marked and removed with Picard (v2.23.8) (https://broadinstitute.github.io/picard/).

Table 2.4. Complete reference genomes

Accession	Description	Study chapter
NC_004337.2	*S. flexneri* 2a, 301 strain, chromosome	3, 4, 5 and 6
NC_004851.1	*S. flexneri* 2a, 301 strain, plasmid	3, 4, 5 and 6
HE616528.1	*S. sonnei*, 53G strain, chromosome	4, 5 and 6
HE616529.1	*S. sonnei*, 53G strain, plasmid A	4, 5 and 6
HE616530.1	*S. sonnei*, 53G strain, plasmid B	4, 5 and 6
HE616531.1	*S. sonnei*, 53G strain, plasmid C	4, 5 and 6
HE616532.1	*S. sonnei*, 53G strain, plasmid E	4, 5 and 6

Variant calling in the mapped reads was performed with samtools mpileup (v1.9) and bcftools (v1.9) against the same reference genome as the reads were mapped to (Table 2.2) [142-144]. A consensus sequence was then defined using the samtools vcfutils.pl, with minimum depth set to 4. Regions of the genome which were likely to have high levels of recombination (plasmids, mobile genetic elements, and phaster (https://phaster.ca/) identified phage sequences) were masked with bedtools (v2.27.1) [142-147]. Phaster was run using the fasta format reference genome sequence.

Consensus sequences, from either pipeline, were compiled into a multiple sequence alignment and run through gubbins (v2.4.1, filter threshold: 31% missing data in Chapters 3 and 6 and the default

25% for Chapters 4 and 5) to identify the core-SNPs, excluding those that likely arose from recombination rather than point mutation [148].

2.5. Genome assembly and annotation

Draft genomes for all studies were assembled using unicycler (v0.4.7) and quality trimmed sequence reads [149]. Genome annotation was achieved with Prokka (v1.14.5) against the *Escherichia* database. Genome assemblies were used to identify genotypes of interest.

2.6. Antimicrobial resistance profiling

In silico identification of genotypic, and prediction of phenotypic, antimicrobial resistance profiles were determined using starAMR (v0.5.1) in Chapter 3, while an optimised profiling method, using both starAMR and AMRfinderPlus (v3.2.3), were used in Chapters 4, 5 and 6. For a selection of antimicrobials for which partial phenotype data was available phenotype prediction for untested isolates was optimised based on strength of association between genotype and phenotype, detailed below. For untested antimicrobials, prediction was made using both AMR genotyping software. Where the same gene was identified but a different phenotype predicted the phenotype predicted by starAMR was used.

All resistance profiling was performed using draft genome assemblies (Section 2.5). Graphic visualisation of resistance profiles, genotypic and phenotypic, was done in R (v4.1.3) using the UpSetR (v1.4.0) package. Profiles were also visualised in the context of population structure (Section 2.7).

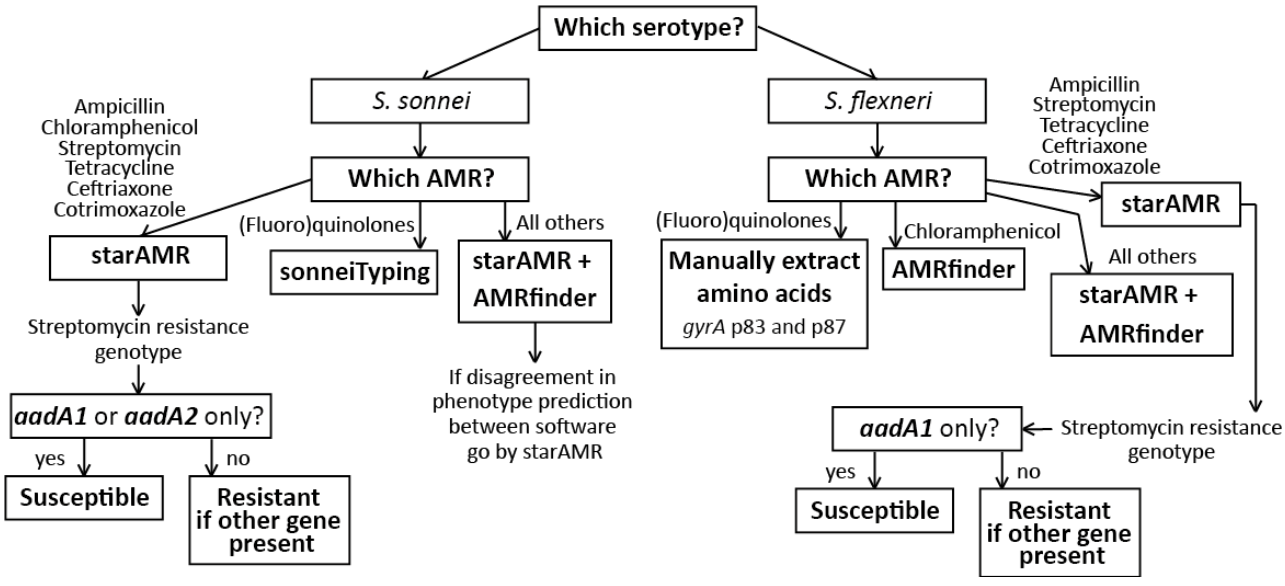

Figure 2.1. AMR phenotype prediction methods decision points.
Shows when and what decisions were made. Starting with choice of genotyping method followed by optimised phenotype prediction acceptance.

2.6.1. Antimicrobial resistance phenotype prediction methods optimisation

A high proportion of the South African study isolates included in Chapters 4 and 5 were phenotypically tested against eight antimicrobials as part of the surveillance program (Table 2.5). Optimisation of the phenotypic resistance prediction accuracy of untested isolates was achieved by comparison (Section 2.6.1.3) of the laboratory determined phenotype (Section 2.6.1.1) against the *in silico* determined genotype and predicted phenotype (Section 2.6.1.2) for these eight antimicrobials (Figure 2.2).

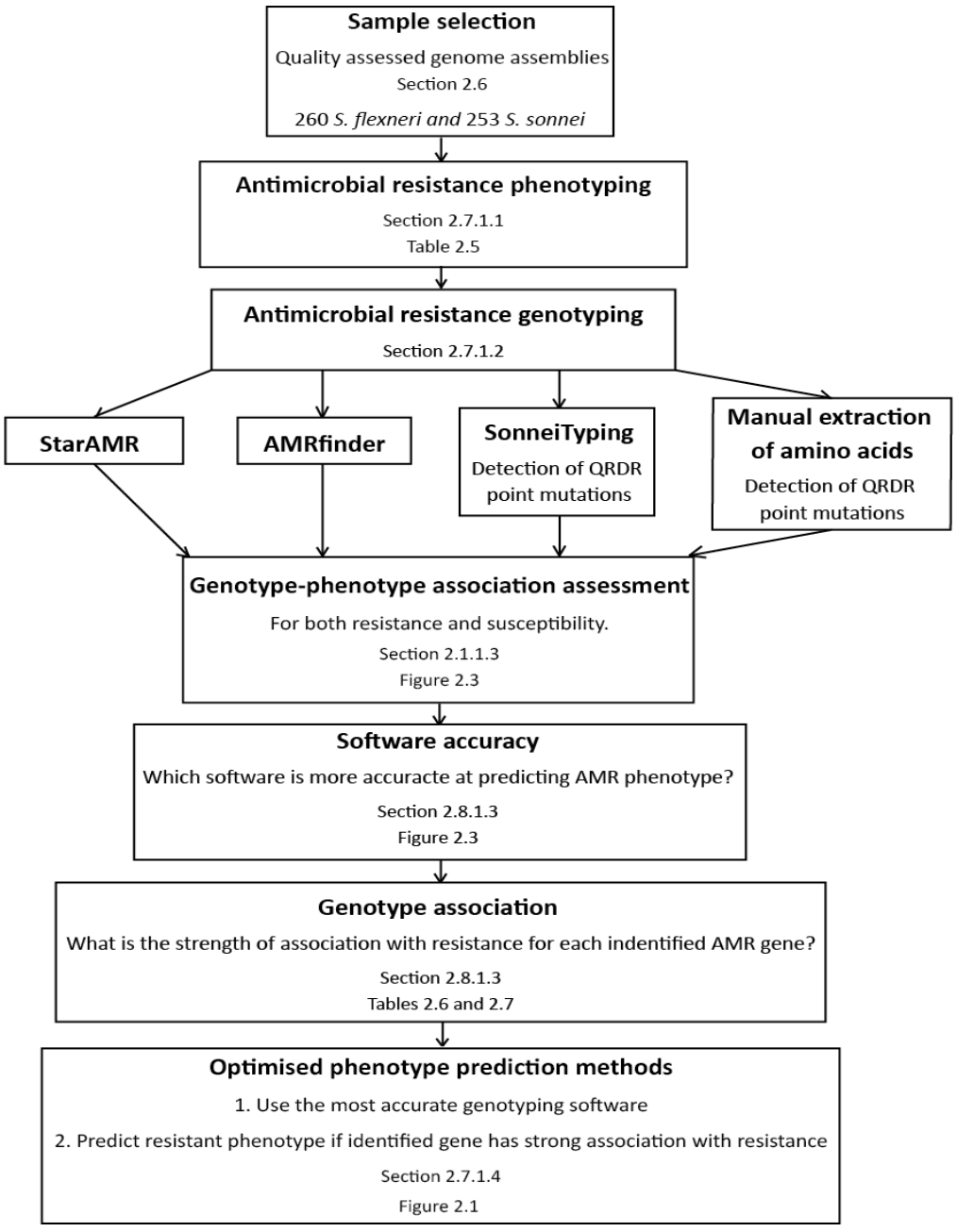

Figure 2.2. Optimisation of antimicrobial resistance phenotype prediction methods for the eight partially phenotype tested antimicrobials.

The optimised methods (detailed in Section 2.6.1.4) were applied only in Chapters 4, 5 and 6, to isolates for which phenotype had not been tested in the laboratory. Where recorded, the laboratory tested phenotype was always accepted as correct. The only isolates to have been phenotype tested in a laboratory were the South African isolates from Chapters 4 and 5.

For untested antimicrobials, genotype and predicted phenotype were identified *in silico* using both starAMR and AMRfinderPlus.

2.6.1.1. *Antimicrobial resistance phenotyping*

Phenotype testing, against ampicillin, chloramphenicol, streptomycin, tetracycline, cotrimoxazole, ceftriaxone, nalidixic acid and ciprofloxacin, was performed on between 64.4% and 95.7% of isolates depending on the antimicrobial class (Table 2.5) as part of routine public health surveillance in South Africa according to standard laboratory procedures (Supplementary) [150-154].

Table 2.5. Number of isolates by antimicrobial resistance phenotype.
Percentage phenotyped = of total isolates, percentage resistant/susceptible = of phenotyped total.

	***S. flexneri* 2a** (n = 260)			***S. sonnei*** (n = 253)		
Antimicrobial	**Phenotyped** n (%)	**Resistant** n (%)	**Susceptible** n (%)	**Phenotyped** n (%)	**Resistant** n (%)	**Susceptible** n (%)
Ampicillin	217 (83.3)	166 (76.5)	51 (23.5)	215 (85.0)	46 (21.4)	169 (78.6)
Chloramphenicol	168 (64.6)	111 (66.1)	57 (33.9)	162 (64.0)	9 (5.5)	153 (94.4)
Streptomycin	168 (64.6)	127 (75.6)	41 (24.4)	163 (64.4)	134 (82.2)	29 (17.8)
Tetracycline	168 (64.6)	115 (68.5)	53 (31.5)	162 (64.0)	134 (82.7)	28 (17.3)
Ceftriaxone	168 (64.6)	0 (0)	168 (100)	162 (64.0)	2 (1.2)	160 (98.8)
Cotrimoxazole	168 (64.6)	124 (73.8)	44 (26.2)	162 (64.0)	149 (92.0)	13 (8.0)
Nalidixic acid	168 (64.6)	0 (0)	168 (100)	162 (64.0)	4 (2.5)	158 (97.5)
Ciprofloxacin	249 (95.7)	1 (0.4)	248 (99.6)	236 (92.9)	0 (0)	236 (100)

The proportion of tested isolates was evenly distributed across South Africa, based on the mean proportion of phenotyped to non-phenotyped isolates from each province across the tested

antimicrobials. Phenotypically detected resistance to chloramphenicol was low in South African *S. sonnei,* while resistance to ceftriaxone, nalidixic acid and ciprofloxacin, was low in both *S. flexneri* 2a and *S. sonnei,* in South Africa (Table 2.5).

2.6.1.2. *Antimicrobial resistance genotyping*

Two AMR genotyping programs were used to detect antimicrobial resistance genes and predict resistance phenotype; starAMR (v0.5.1) (https://github.com/phac-nml/staramr) and AMRfinderPlus (v3.2.3) [155]. Each software determines AMR genotype by *in silico* comparison of genome assemblies against an AMR gene database. For starAMR, this is the commonly used resfinder database while for AMRfinderPlus this is the newer NCBI Bacterial Antimicrobial Resistance Reference Gene Database (https://www.ncbi.nlm.nih.gov/bioproject/PRJNA313047) [155, 156].

Genes in both databases are linked with an associated resistance phenotype which is reported alongside the gene name within the program output and was taken as a resistance phenotype prediction. For both pieces of software, gene presence, and predicted phenotype, were accepted when isolate sequence identity was ≥99% of the database sequence and coverage ≥90%.

While starAMR detects only gene mediated resistance, AMRfinderPlus detects some point mutations such as those in the quinolone resistance determining region (QRDR). The *S. sonnei* typing software sonneiTyping sonnei_genotype.py (v1) (https://github.com/katholt/sonneityping) also detects point mutations in the QRDR, the accuracy of which was assessed alongside those detected by AMRfinderPlus [39]. The accuracy of the detected QRDR point mutations was assessed against quinolone resistance phenotype (resistance to nalidixic acid and/or ciprofloxacin) and manually determined genotype. Manual determination of point mutations involved the *in silico* extraction of amino acids at known resistance associated sites in the QRDR, *gyrA* (positions 83 and 87) and *parC* (positions 80 and 91), from annotated draft genomes [78, 157].

2.6.1.3. *Antimicrobial resistance genotype-phenotype comparison*

Comparing software phenotype prediction accuracy

To determine which genotyping software would provide a more accurate phenotype prediction I compared the proportion of correctly predicted resistant and correctly predicted susceptible isolates by each software.

The same levels of predicted resistance were produced by both software for ampicillin, chloramphenicol and ceftriaxone in *S. sonnei*, and ampicillin and cotrimoxazole in *S. flexneri* 2a (Figure 2.3). Where a discrepancy between the two software was observed, starAMR was generally found to be more accurate. AMRfinderPlus was only more accurate at predicting chloramphenicol phenotype, in *S. flexneri,* correctly identifying 93% of the susceptible isolates (susceptible n = 57) vs 91% by starAMR; both software were equally accurate at predicting resistance (Figure 2.3A).

AMRfinderPlus was overall found to be worse at predicting streptomycin phenotype in *S. sonnei,* despite being slightly more accurate at predicting resistance (100% vs 99% of 127 resistant isolates), as it far less accurate than starAMR at predicting susceptibility (10% vs 24% of 41 susceptible isolates) (Figure 2.3B). Both software were poor predictors of streptomycin resistance, identifying streptomycin resistance associated genes in a high proportion of the phenotypically susceptible isolates. Given the stringency of the gene identity thresholds used to accept functional gene presence it is likely that the susceptible phenotype of these isolates is due to gene silencing rather than erroneous detection of absent genes.

No quinolone resistance associated genes were identified in the South African dataset, all resistance was likely conferred through point mutations in the QRDR. The QRDR point mutations detected by AMRfinderPlus were found to be inaccurate when compared against both the manually extracted amino acids at the sites corresponding to those detected by the software (inaccurate amino acid identification) and the recorded phenotype (inaccurate phenotype prediction), for both serotypes.

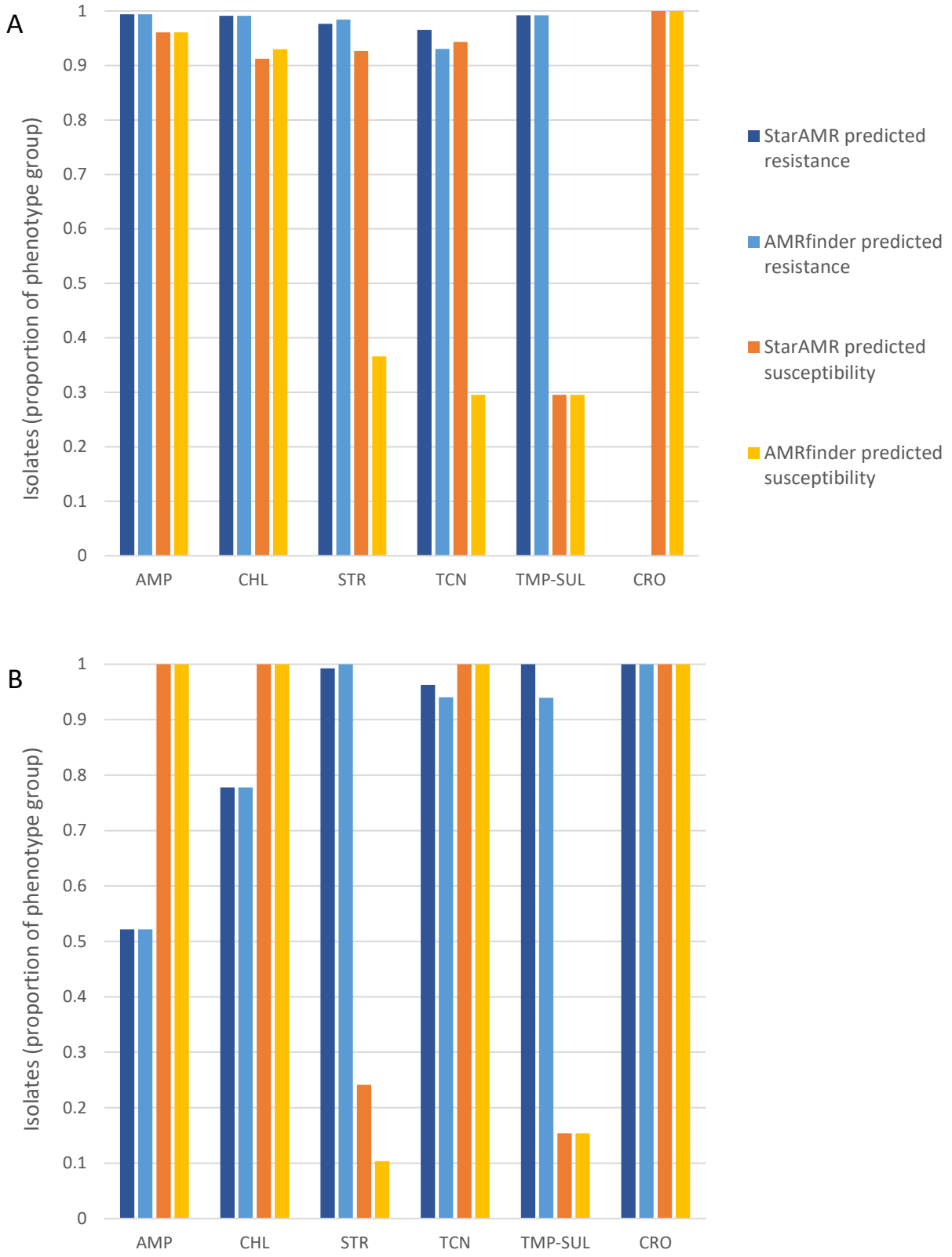

Figure 2.3. Accuracy of AMRfinderPlus and StarAMR AMR genotyping software at predicting AMR resistance and susceptibility among South African S. flexneri 2a (A) and S. sonnei (B).

Total number of isolates varies by antimicrobial. AMP = ampicillin, CHL = chloramphenicol, STR = streptomycin, TCN = tetracycline, TMP-SUL = cotrimoxazole, CRO = ceftriaxone.

The sonneiTyping QRDR point mutations detection was found to be accurate (for *S. sonnei*) against the manually extracted amino acids. In most cases (4/5), these detected point mutations correlated with a resistant phenotype, the only exception was *gyrA* D87Y which had a 50% association with resistance (1/2 isolates) to nalidixic acid. No other resistance determinant was identified in the discrepant isolate which could explain the difference, this might be due to human error in the phenotype recording.

Based on the accuracy comparison, I decided to use starAMR to detect genotype and predict phenotype for all eight tested antimicrobials for *S. sonnei* and all but chloramphenicol for *S. flexneri*. I instead used AMRfinderPlus to predict chloramphenicol resistance phenotype in *S. flexneri*. The *S. sonnei* typing software sonneiTyping software was used to detect resistance associated QRDR point mutations and predict quinolone resistance in *S. sonnei*, while manual extraction of amino acids was relied upon for *S. flexneri*.

Genotype-phenotype comparison

The level of agreement between genotype/predicted phenotype and tested phenotype was also examined for each individual resistance determinant, to exclude any which were poor predictors of resistance. This was done in two ways: 1) the percentage resistance in total isolates with gene, 2) the percentage resistance in isolates where gene was only resistance determinant present (for the relevant antimicrobial).

Most genes were good predictors of resistance though a few were poor (Tables 2.6 and 2.7). In *S. sonnei,* both *aadA1* and *aadA2* alone were a poor predictors of streptomycin resistance, 40% of isolates with only *aadA1* and no other streptomycin resistance gene were phenotypically streptomycin resistant, while zero isolates with *aadA2* alone had a resistant phenotype (Table 2.7). Based on the poor association with resistance, streptomycin resistance was not predicted when either *aadA1* or *aadA2* were the only streptomycin resistance determinant present.

Table 2.6. Number of antimicrobial resistance phenotyped S. flexneri 2a isolates per identified resistance gene and the proportion of isolates with an identified resistance gene with a resistant phenotype.
Showing both the total number of resistant isolates identified with the resistance gene as well as the number of resistant isolates where the resistance gene is the only gene present conferring resistance to that antimicrobial. * = predicted by AMRfinderPlus rather than StarAMR.

Antimicrobial	**Resistance gene**	**Isolates – all**		**Isolates – single gene**	
		n	Resistant phenotype n (%)	n	Resistant phenotype n (%)
Ampicillin	*blaOXA-1*	153	151 (99)	150	148 (99)
	blaTEM-1A	1	1 (100)	1	1 (100)
	blaTEM-1B	16	16 (100)	13	13 (100)
Chloramphenicol *	*catA1*	114	110 (96)	114	110 (96)
Streptomycin	*aadA1*	121	119 (98)	56	54 (96)
	aph(3'')-Ib	71	70 (99)	6	5 (83)
Tetracycline	*tet(A)*	7	6 (86)	5	4 (80)
	tet(B)	109	107 (98)	107	105 (98)
Co-trimoxazole (Trimethoprim)	*dfrA1*	121	91 (75)	111	81 (73)
	dfrA5	2	2 (100)	2	2 (100)
	dfrA7	1	1 (100)	-	-
	dfrA8	5	5 (100)	2	2 (100)
	dfrA14	35	34 (97)	27	26 (96)
	dfrA15	1	1 (100)	1	1 (100)
(Sulfisoxazole)	*sul1*	3	3 (100)	1	1 (100)
	sul2	106	104 (98)	104	102 (98)

Quinolone resistance associated point mutations in the QRDR were detected in both serotypes. The D87Y *gyrA* point mutation (2 isolates) had only a 50% association, perhaps due to human error in the phenotyping. Due to evidence from the literature, and the 50% association with resistance found in the phenotyped isolates, this determinant was still taken as resistance associated. No resistance

determinant was identified in the only phenotypically ciprofloxacin resistant *S. flexneri* isolate, this may also be due to human error in phenotyping, or it could be due to an undetected determinant.

Table 2.7. Number of antimicrobial resistance phenotyped S. sonnei isolates per identified resistance gene and the proportion of isolates with an identified resistance gene with a resistant phenotype.
Showing both the total isolates identified with the resistance gene and isolates where the resistance gene is the only gene present conferring resistance to that antimicrobial. ^ = intermediate resistance.

Antimicrobial	**Resistance gene**	**Isolates – all**		**Isolates – single gene**	
		n	Resistant phenotype n (%)	n	Resistant phenotype n (%)
Ampicillin and ceftriaxone	*blaCMY-2*	1	1 (100)	1	1 (100)
	blaCMY-4	1	1 (100)	1	1 (100)
Ampicillin	*blaOXA-1*	7	7 (100)	7	7 (100)
	blaTEM-1B	17	17 (100)	17	17 (100)
Chloramphenicol	*catA1*	6	6 (100)	6	6 (100)
	*cmlA1^*	1	1 (100)	1	1 (100)
Streptomycin	*aadA1*	153	131 (86)	35	14 (40)
	aadA2	1	0 (0)	-	-
	aph(3'')-Ib	119	119 (100)	8	8 (100)
Tetracycline	*tet(A)*	123	123 (100)	121	121 (100)
	tet(B)	8	8 (100)	6	6 (100)
Co-trimoxazole (Trimethoprim)	*dfrA1*	154	144 (94)	135	125 (93)
	dfrA7	1	1 (100)	-	-
	dfrA8	2	2 (100)	-	-
	dfrA12	1	1 (100)	1	1 (100)
	dfrA14	20	20 (100)	3	3 (100)
	dfrA15	1	1 (100)	-	-
(Sulfisoxazole)	*sul1*	1	1 (100)	-	-
	sul2	141	140 (99)	140	139 (99)
	sul3	2	2 (100)	1	1 (100)

A large under prediction of resistance was seen for ampicillin and chloramphenicol in *S. sonnei* (Figure 2B). This suggests one or more undetected resistance determinants are likely present in the population. There are several possible causes for a lack of detection. As no detected ampicillin or chloramphenicol resistance genes were excluded due to quality thresholds it was not possible to make threshold adjustments to improve the prediction accuracy. Other possible explanations include gene interruption in the genome assembly, AMR determinant on an unstable plasmid, incomplete database, or poor sequence or assembly quality. No further work was done to try and identify these undetected resistance determinants as this fell outside the scope of this project.

2.6.1.4. *Optimised methods*

For the eight partially phenotyped antimicrobials the resistance phenotype of untested isolates was predicted based on genotype. This prediction was optimised based on the results laid out above and involved 1) starAMR to predict phenotype in both serotypes, except for chloramphenicol in *S. flexneri*, 2) rejecting a streptomycin resistant phenotype prediction if only *aadA2* present, 3) rejecting a streptomycin resistant phenotype prediction, in *S. sonnei,* if only *aadA1* present, 4) manual extraction of amino acids resistance associated QRDR sites for *S. flexneri,* 5) the use of sonneiTyping software to predict quinolone resistance based on QRDR point mutations.

2.7. Phylogenetic tree visualisation

All phylogenies were visualised using the Interactive Tree of Life (iTOL) (v6.5.7) web-based tool, typically alongside associated isolate/patient metadata [158]. Branch lengths, node distances, and bootstrap/support values were all determined through the FigTree (v1.4.4) phylogeny visualisation software (http://tree.bio.ed.ac.uk/software/figtree/).

Chapter 3

Whole genome sequence analysis of *Shigella* from Malawi

Preface

This chapter examines the *Shigella* isolates collected at a single hospital from children under the age of five, 2012 to 2015, and includes all identified *shigellae* regardless of serotype.

The content of this chapter was published under the title "Whole genome sequence analysis of *Shigella* from Malawi identifies fluoroquinolone resistance" in Microbial Genomics [159]. The contribution of my co-authors to the experimental work in this chapter is acknowledged below, all other work was completed by myself.

Khuzwayo C. Jere	Assisted with the collection of isolates from MLW.
Chikondi Peno	Assisted with the collection of isolates from MLW.
Rebecca J. Bengtsson	Provided bioinformatic support.
End Chinyama	Assisted with the collection of isolates from MLW.
Jonathan Mandolo	Assisted with the collection of isolates from MLW.
Amy K. Cain	Assisted with the collection of isolates from MLW.
Naor Bar-Zeev	Assisted with the collection of isolates from MLW.
Nigel Cunliffe	Sequencing of isolates.
Jennifer Cornick	Conceptualization and access to isolates.

3.1. Introduction

Shigella is the second leading cause of diarrhoeal death globally with the greatest disease burdens seen in low- and middle-income countries (1). Approximately one third of these deaths are in children under the age of five years old, with infection potentially causing chronic health effects such as stunted growth, reduced cognitive development, and chronic, functional bowel disorders (1-9). Growing antimicrobial resistance (AMR) is increasingly limiting treatment options and threatens to reverse hard won reductions in diarrhoeal mortality in high-burden areas (10). Vaccination will likely be a potential solution in the future; however, many candidates are still in development. Vaccine candidates can be split into two groups 1) those targeting immunogenic surface molecules, replicating natural immunity, which is serotype specific, and 2) those targeting conserved protein epitopes, aiming to be more broadly protective (11). The efficacy of both approaches may be limited by *Shigella* diversity. Characterisation of antimicrobial resistance determinants is important to ensure ongoing effectiveness of treatment regimens, while characterisation of strains is important for ensuring appropriate vaccine candidate choice.

Fluoroquinolone resistant (FQR) strains are of particular importance as there are few widely effective alternatives to this, currently, first-line treatment. The importance of FQR strains is reflected by their inclusion on the WHO global priority pathogens list (12). Fluoroquinolone resistance can be acquired *de novo,* through a double mutation in the *gyrA* gene (amino acids 83 and 87) of the quinolone resistance determining region (QRDR), with a third mutation in the *parC* gene (AA80) ameliorating the fitness cost, or through horizontal transmission of FQR genes such as the *qnr* genes (13). Mutations in *gyrA* act to disrupt the ability of the fluoroquinolone molecules binding to the DNA gyrase enzyme product, inhibiting fluoroquinolone function. DNA topoisomerase IV, encoded by *parC*, can be a secondary target of fluroroquinolones. The binding of the *qnr* gene products to these enzymes likewise blocks the fluoroquinolone function (14).

Shigella is reported as a leading cause of diarrhoea among hospitalised children in Malawi, however, there is little information on the circulating strains (15). Whole genome sequence analysis (WGSA) has been successfully applied to investigate *Shigella* epidemiology and AMR determinants and can greatly aid in disease control in high-burden areas such as Malawi (16). Here, I applied WGSA to characterise *Shigella* strains and AMR determinants in Malawi. Providing important baseline information for public health interventions including antibiotic treatment and deployment of *Shigella* vaccines, the development of which is a WHO priority (17).

3.2. Methods

3.2.1. Sample selection and sequencing

All ten biochemically identified shigellae collected during a rotavirus vaccine evaluation programme, between 2012 and 2015, were included in this study. All were isolated from faecal samples taken from children under the age of five years old, hospitalised with acute gastroenteritis at the Queen Elizabeth Central Hospital, Blantyre, Malawi, and subjected to whole genome sequencing at the Wellcome Trust Sanger institute according to in-house protocols (18, 19, 20).

3.2.2. Quality control

Read sequence quality trimming was performed as laid out in the methods chapter, 3.7 quality control section. All unpaired reverse reads from the Malawi isolates were excluded and an additional thirty-three bases were removed from the ends of the remaining Malawi reads due to poor quality of these bases post-trimming.

3.2.3. Assembly

Draft genomes were assembled as laid out in the methods chapter, 3.6 genome assembly section, and quality assessed as laid out in the methods chapter, 3.7 quality control section, against the complete *S. flexneri* strain 301 genome (chromosome and plasmid, NC_004337.2 and NC_004851.1 respectively) (22, 23).

3.2.4. Species confirmation

Species were confirmed with a combination of in silico multi-locus sequence typing (MLST; https://github.com/tseemann/mlst), BLAST (v2.10.0) searches of the isolate draft genomes against the nucleotide database, and generation of a maximum likelihood phylogeny of the *Shigella*/*Escherichia coli* clade (Figure 3.1) (21). The MLST sequence type (ST), generated from the draft genomes, was compared against the enterobase database (http://enterobase.warwick.ac.uk/) to identify species with the same sequence type. Only those identified as *Shigella* or *E. coli* (8/10), using MLST ST supported by the BLAST search results, were included in the study. Blast identification was

done by identifying species of origin for BLAST hits (e-value = 0.0) against the first ten draft genome contiguous sequences (contigs), taking the most frequently identified species. Two isolates were excluded on this basis, identified as *Escherichia albertii* and *Providencia alcalifaciens* (Table 3.1). All *Shigella* isolates were serotyped in silico with ShigaTyper (v1.0.6, https://github.com/CFSAN-Biostatistics/shigatyper).

Table 3.1. MLST, BLAST comparison, and shigaTyper results.

Isolate ID	MLST - ST	MLST	BLAST	ShigaTyper
22204_7#73	4619	*Escherichia albertii*	*Escherichia albertii*	EIEC
22204_7#74	145	*S. boydii / flexneri*	*S. boydii*	*S. boydii* 2
22204_7#75	-	-	*Providencia alcalifaciens*	Not *Shigella* or *E. coli*
22204_7#76	8221	*S. boydii*	*Escherichia coli*	EIEC
22204_7#77	6	*Escherichia coli*	*Escherichia coli / S. boydii*	EIEC
22204_7#78	630	*S. flexneri*	*S. flexneri*	*S. flexneri* 4av
22204_7#79	145	*S. boydii / flexneri*	*S. flexneri / boydii*	*S. flexneri* 6
22204_7#80	145	*S. boydii / flexneri*	*S. flexneri / boydii*	*S. flexneri* 6
22204_7#81	245	*S. flexneri*	*S. flexneri*	*S. flexneri* 3a
22204_7#82	245	*S. flexneri*	*S. flexneri*	*S. flexneri* 3a

3.2.5. Maximum likelihood phylogeny

The maximum likelihood phylogeny was generated as laid out in the methods chapter, 3.5 maximum-likelihood phylogenetics section. The generated core-SNP alignment (40075 SNPs) included all study isolates identified as *Shigella* or *E. coli* and a selection of reference isolates chosen from across the known *E. coli*/*Shigella* phylogeny to enable tree reconstruction (Table 2.1) (24).

3.2.6. Identification of enteroinvasive *E. coli*

To determine if the identified *E. coli* isolates were EIEC, I looked for the presence of the mxi-spa locus, found on the large virulence plasmid (pINV). Contiguous sequences (contigs) from quality-assessed isolate draft-genomes were compared against a reference EIEC pINV (CP011417.1), with BLAST (v2.10.0), to identify isolate contigs possibly encoding the mxi-spa locus. These contigs were then visually compared against the reference pINV, using the Artemis Comparison Tool (33), to confirm a match against the mxi-spa locus.

3.2.7. Antimicrobial resistance genotyping

Genotypic AMR profiles were determined for *Shigella* isolates as laid out in the methods chapter, 3.9 antimicrobial resistance profiling section, with the exception of using starAMR for chloramphenicol resistance rather than AMRfinder. Predicted resistance phenotype was extracted directly from the genotyping software. Point mutation mediated FQR was identified by comparing the *gyrA* and *parC* amino acid sequences across all isolates, extracted from the prokka (v1.14.5) annotated draft genomes (38), with amino acid identity at resistance-associated sites confirmed as expected based on the literature (13).

3.2.8. Mobile genetic element identification

To identify possible plasmids, all isolate assembly contigs were compared against the plasmidfinder database (35). Only contigs with sequenced identity and overlap of ≥98% and ≥80% respectively were accepted as likely being plasmid contigs. ISfinder was used to identify insertion sequences in AMR gene carrying contigs [160]. Presence of an Inc(F) multi-drug resistant (MDR) plasmid was confirmed through mapping of isolate sequence reads using bwa mem (v0.7.5) to the reference *S. flexneri* 1c strain AUSMDU00008355 plasmid 3 (LR213454) (39).

3.3. Results

From our eight isolates, four *Shigella* serotypes were identified: two *S. flexneri* 3a (Sf3a, Phylogroup 2), one *S. flexneri* 4av (Sf4av, Phylogroup 7), two *S. flexneri* 6 (Sf6) and one *S. boydii* 2 (Sb, Clade 3). Two isolates were *E. coli* (Figure 3.1A), one of which was identified as EIEC (99% identity and 100% coverage with mxi-spa locus).

All *Shigella* isolates were predicted to be MDR, carrying genes conferring resistance to three or more drug classes (Table 3.2). While overall resistance was high, I observed limited diversity in AMR genes and predicted resistance profiles; 13 genes encoding resistance to 7 antimicrobial classes (Figure 3.1B, Table 3.2). One isolate (Sf4av) was predicted FQR, due to the presence of *qnrS1* (Table 3.2). A mutation at *parC* R91Q was also identified, however, the same variation was present in all isolates, *E. coli* and *Shigella*, and likely represents natural variation rather than a resistance adaptation. Unfortunately, phenotype data were unavailable to confirm these findings.

Resistance gene encoding contigs from two isolates were identified as likely plasmid contigs, based on a match with plasmidfinder database. One Sf3 isolate (22204_7#81) carried a MDR IncQ1 plasmid, and the Sf4av isolate carried a *tet(A)* encoding IncFIB(K) plasmid (Table 3.2). Another MDR contig had a read depth 5.29-fold higher than the chromosomal contigs, possible evidence of a multi-copy plasmid (Table 3.2).

The same AMR gene profile (*aph(3'')-Ib, aph(6)-Id* and *sul2*) was identified in contigs from multiple isolates (22204_7#74, 22204_7#78 and 22204_7#79) across distinct phylogroups (Sb2 clade 3, Sf4av Phylogroup 7 and Sf6 respectively) (Figure 3.1, Table 3.2). However, a pairwise BLAST comparison showed that these contigs had similarity (≥99% sequence identity) only across the AMR gene encoding region (2620bp - 3184bp). The AMR gene encoding region of the largest of these contigs (Sf4av) contains insertion sequences and is surrounded by transposase genes, with AMR gene (*blaTEM-1B*) identified as being within a Tn2 transposon (e-score = 0.0 and database sequence coverage = 100%) (Figure 3.2B) (37).

Table 3.2. Antimicrobial resistance genotypic and predicted phenotypic profiles of Malawian Shigella isolates by contiguous sequence.

[ω] StarAMR identified IncFIB(K) plasmid, [ר] StarAMR identified MDR IncQ1 plasmid, [ד] possible multi-copy plasmid.

Isolate ID	Contig length (bp)	Resistance genes	Predicted resistance drug class
22204_7#74	9131 [ω]	*tet(A)*	Tetracycline
(*S. boydii* 2)	2798	*aph(3'')-Ib, aph(6)-Id, sul2*	Aminoglycoside, sulfonamide
	2353	*dfrA7*	Trimethoprim
	1831	*blaTEM-1B*	Aminopenicillin
22204_7#78	48082	*tet(A)*	Tetracycline
(*S. flexneri* 4av)	17334	*aph(3'')-Ib, aph(6)-Id, sul2, blaTEM-1B, qnrS1*	Aminoglycoside, sulfonamide, aminopenicillin, fluoroquinolone
	1588	*dfrA14*	Trimethoprim
22204_7#79	34138	*tet(A), sul1, dfrA7, catA1*	Tetracycline, sulfonamide, trimethoprim, chloramphenicol
(*S. flexneri* 6)	6200	*aph(3'')-Ib, aph(6)-Id, sul2*	Aminoglycoside, sulfonamide
22204_7#80	25733	*tet(A), sul1, dfrA7, catA1*	Tetracycline, sulfonamide, trimethoprim, chloramphenicol
(*S. flexneri* 6)	6200	*aph(3'')-Ib, sul2*	Aminoglycoside, sulfonamide
22204_7#81	11392 [ר]	*dfrA14, sul2, tet(A)*	Trimethoprim, sulfonamide, tetracycline
(*S. flexneri* 3a)			
22204_7#82	45377	*tet(B)*	Tetracycline
(*S. flexneri* 3a)	8773	*aadA1, blaOXA-1, catA1*	Aminoglycoside, aminopenicillin, chloramphenicol
	6790 [ד]	*aph(6)-Id, dfrA14, sul2*	Aminoglycoside, trimethoprim, sulfonamide

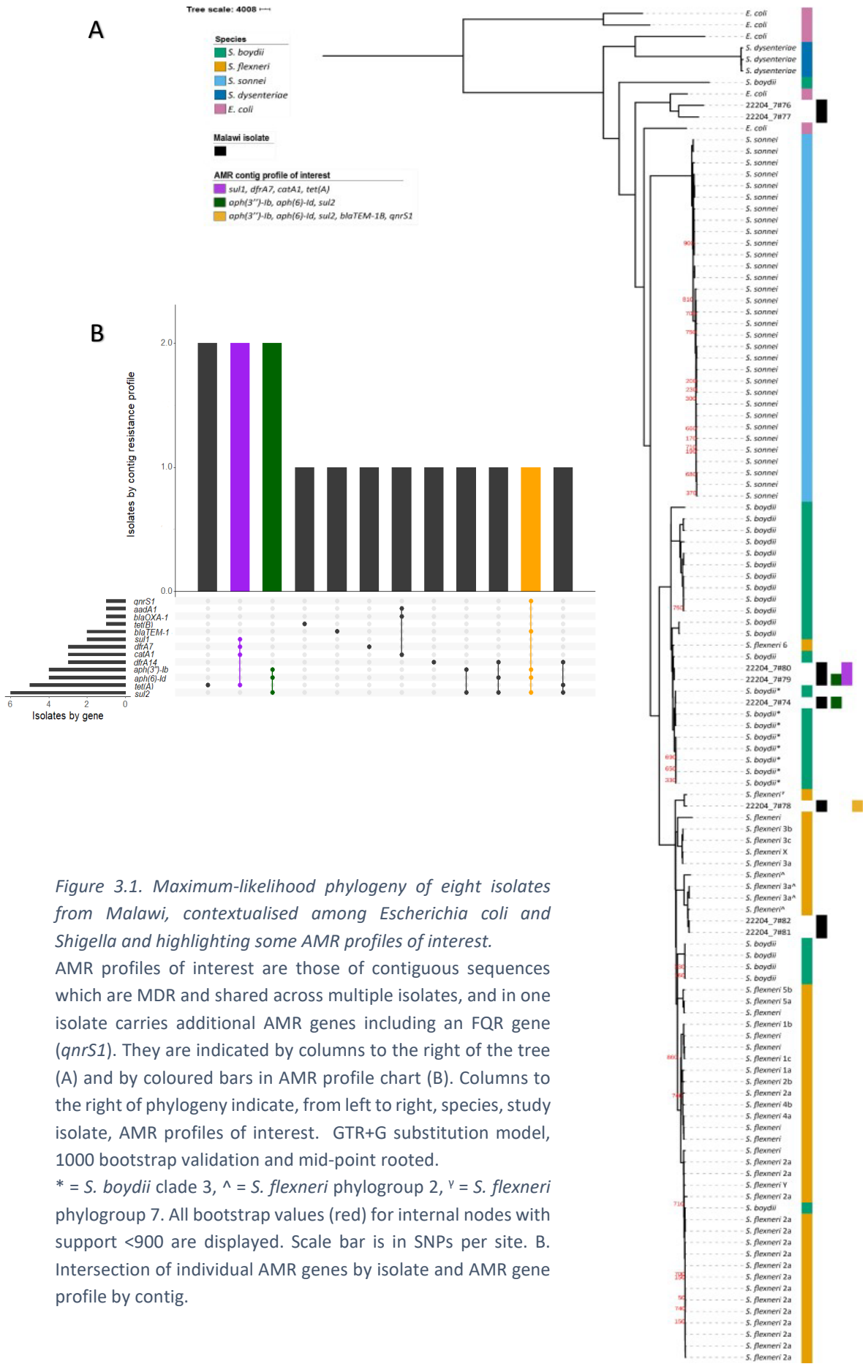

Figure 3.1. Maximum-likelihood phylogeny of eight isolates from Malawi, contextualised among Escherichia coli and Shigella and highlighting some AMR profiles of interest.

AMR profiles of interest are those of contiguous sequences which are MDR and shared across multiple isolates, and in one isolate carries additional AMR genes including an FQR gene (*qnrS1*). They are indicated by columns to the right of the tree (A) and by coloured bars in AMR profile chart (B). Columns to the right of phylogeny indicate, from left to right, species, study isolate, AMR profiles of interest. GTR+G substitution model, 1000 bootstrap validation and mid-point rooted.

* = *S. boydii* clade 3, ^ = *S. flexneri* phylogroup 2, ˠ = *S. flexneri* phylogroup 7. All bootstrap values (red) for internal nodes with support <900 are displayed. Scale bar is in SNPs per site. B. Intersection of individual AMR genes by isolate and AMR gene profile by contig.

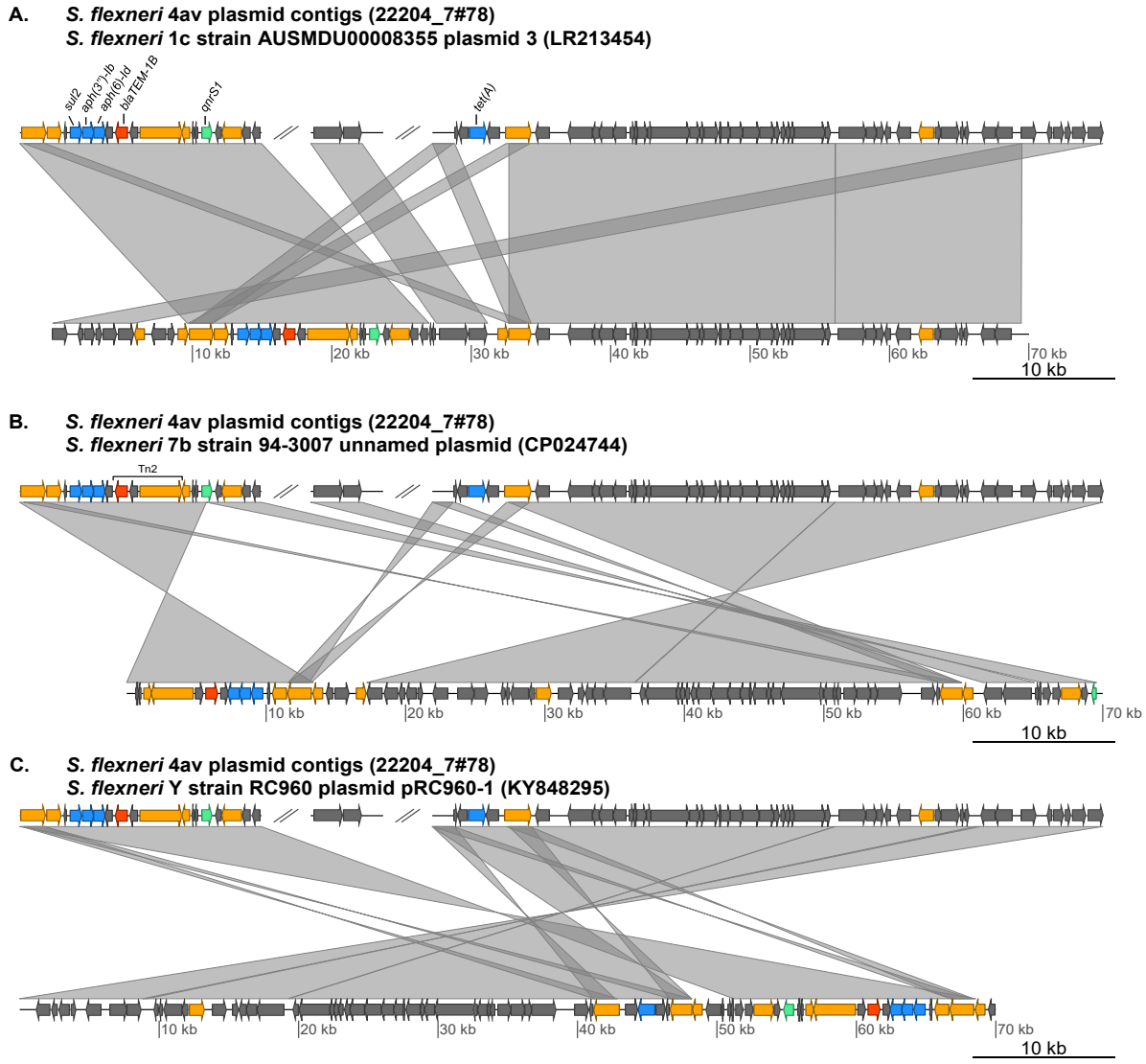

Figure 3.2. Pairwise comparisons of S. flexneri 4av study isolate plasmid contigs against previously identified MDR plasmids.

Pairwise comparisons the *S. flexneri* 4av isolate plasmid contigs (upper sequence) against the previously identified plasmids (lower sequences) generated with GenoPlotR package in R from prokka annotated genomes, compared using BLAST. Transposase genes are shown in orange, AMR genes are in blue, except FQR gene *qnrS1* which is in green and β-lactamase gene *blaTEM-1B* which is in red and is part of a Tn2 transposon. **A.** Comparison against the *S. flexneri* 1c, strain AUSMDU00008355 plasmid 3, which lacks a *tet(A)* gene. **B.** Comparison against the *S. flexneri* 7b, strain 94-3007 unnamed plasmid, which also lacks a *tet(A)* gene. **C.** Comparison against the *S. flexneri* Y, strain RC690 plasmid pRC960-1.

The sf4av *tet(A)* encoding IncFIB(K) plasmid was found to likely be part of a MDR plasmid in combination with other contigs (Figure 3.2). A BLAST comparison of the *qnrS1*-containing contig (17334 bases) against the NCBI nt database identified three top equivocal hits against *S. flexneri* plasmids (KY848295.1, CP024474.1, LR213454.1) (e-score = 0.0, query sequence coverage and identity

≥99.9%). BLAST comparison of the isolate draft genome contigs against all three of these plasmids identified the same three contigs with hits against all three plasmids (e-value = 0.0 and bit score ≥ 999). Pairwise comparisons of these contigs showed high sequence similarity and sequence coverage against all three plasmids, which were themselves highly similar to each other (Figure 3.2). Read mapping coverage (mean depth = 53 reads) provides further evidence that the isolate was carrying a plasmid very similar to these other previously identified plasmids. All four of the plasmids are MDR, encoding *sul2, aph(3'')-Ib, aph(6)-Id, blaTEM-1B* and *qnrS1* genes. Our isolate and one other also carried an additional resistance gene *tet(A)* (Figure 3.2). Each was identified in a different *S. flexneri* serotype (1c, Y, 7b and 4av).

3.4. Discussion

The lack of *S. sonnei* isolates in our collection was expected as *S. sonnei* (Ss), though highly prevalent globally, is typically associated with high-income nations and industrialisation (28). Prevalence of *S. boydii* in Africa is thought to be low, isolated in 5.7% of cases of a multi-national study (29). However, detection of this serogroup in this study would still be likely by chance, even with the small sample size (binomial test p-value = 0.375), and may, therefore, not be indicative of higher prevalence in Malawi. The strain diversity shows multiple, distinct *Shigella* strains circulate in Malawi; further characterisation is needed to aid effective vaccine development for the region. Particularly as a 2016 *Shigella* vaccine candidate review recommended multivalent vaccines target Sf2a, Sf3a, Sf6 and Ss, being the most prevalent globally, which would fail to protect against all the strains identified in this study (11).

Identification of *E. coli* among the samples is likely due to the close relatedness between the two species, as *Shigella* is a specialised pathovar of *E. coli* and shares a disease phenotype with enteroinvasive *E. coli* (EIEC) mediated by the large virulence plasmid pINV. A plasmid carrying the mxi-spa locus which encodes a Type 3 Secretion System and secreted effector proteins (30-32).

Antimicrobial resistance is a growing issue globally and my results demonstrate MDR *Shigella* strains circulate in Malawi. The limited diversity suggests that there are treatments which remain effective in Malawi, such as azithromycin, though this would need confirming in a larger study and may change with mass drug administration programs.

Mobile genetic elements (MGE) have been shown to be important drivers of AMR dissemination globally (36). The identification of isolates which likely possess AMR carrying plasmids and other MGE support the role of MGE in the spread of AMR in Malawi. Of note, the detection of a *qnrS1* carrying plasmid shows that acquirable FQR is present among *Shigella* in Malawi. There is, therefore, a high risk of widespread FQR in Malawi and neighbouring regions, though further study into the prevalence and nature of FQR in the region is needed. The MDR plasmids similar to the identified *qnrS1*-carrying

plasmid, were from multiple *S. flexneri* serotypes, across multiple lineages and continents (specifically America (unpublished), Australia, China, and Malawi) (40-42). This suggests the MDR plasmid has been horizontally transmitted among strains and spread intercontinentally, providing further evidence for an interaction between MDR *Shigella* globally and MDR *Shigella* in Malawi, though further research is needed to characterise this.

3.4.1. Conclusions

Together, the high proportion of MDR *Shigella,* the acquirable FQR and the MDR plasmids detected in this study show that, without intervention, controlling shigellosis in Malawi will be increasingly difficult, which will likely have global consequences. Meanwhile, the diversity of strains identified in this study indicate choice of vaccine candidate in Malawi will need to be selected based on local, rather than global, *Shigella* strain prevalence. Highlighting the importance of further research into the epidemiology of *Shigella* in the region in ensuring effective disease control.

Chapter 4

Genomic Epidemiology of shigellosis in South Africa, part 1: *Shigella flexneri* 2a

Preface

This chapter forms one half of the national level study, which used a sub-sample of isolates collected as part of the ongoing public healthcare surveillance in South Africa, limited to the two most prevalent serotypes. This chapter focuses on *S. flexneri* 2a, the most prevalent serotype across the study period. This work was aided by the work of several collaborators whose contributions are outlined in the table below.

Anthony Smith	Provided South African isolate metadata
Juno Thomas	Provided South African isolate metadata, official level of urbanisation classifications for South African provinces, and information on public healthcare policy and processes in South Africa
Karen Keddy	South African isolates sample selection Background South African public health information.
Neil Hall	Whole genome sequencing
Rebecca J. Bengtsson	Provided accession numbers for the *S. flexneri* phylogenetic reference isolates, having created known global population structure-representative, serogroup isolate lists, and smaller serogroup representative isolate lists. Also provided a working core-SNP alignment generation pipeline and bioinformatics support.

4.1. Introduction

Shigella flexneri is an important global serogroup which is prevalent across the globe but dominant in developing regions [35, 161]. The serogroup is old, most recent common ancestor (MRCA) est. 1848, and highly diverse; seven Phylogroups within the global population have been identified to date, each typically contains multiple serotypes and is globally distributed [33]. This level of global diversity is likely a consequence of the ability of multiple strains of *S. flexneri* to co-exist in the same region, rather than undergoing strain replacement [33].

Serotype 2a is the most prevalent serotype in South Africa, accounting for nearly 40% of cases [150-152]. The serotype has been predominantly found in Phylogroup 3, a phylogroup associated with the *Shigella* resistance locus 1 (SHI-1) [33]. Its population is predicted to be predominantly MDR, however a predicted drug susceptible sub-population of this phylogroup has also been found [33].

Shigellosis in South Africa is included as a surveillance pathogen by the Group for Enteric, Respiratory and Meningeal Diseases Surveillance in South Africa (GERMS-SA). The surveillance program involves the collection and storage of isolates from public healthcare hospitals from across all nine provinces [150-154]. From the annually released reports, *S. flexneri* 2a and *S. sonnei* are the most common serotypes, with their dominance increasing with time from 2009. Together they accounted for nearly 70% of cases from 2011 to 2014 [150-154, 162-164]. Other *S. flexneri* serotypes (1b, 3a, 6) were the cause of around 24% of other cases.

Resistance data in the GERMS-SA reports is incomplete, particularly after 2015, however, what data is available shows intermediate to high levels of resistance to ampicillin, co-trimoxazole, tetracycline, streptomycin, and chloramphenicol and susceptibility to nalidixic acid, ciprofloxacin and ceftriaxone [150-154, 164].

In this study I used WGSA to examine the genomic epidemiology of *S. flexneri* 2a in South Africa, to describe the population structure using phylogenetics, model population dynamics, and analyse the

potential influences of AMR and virulence characteristics of the identified strains on the observed epidemiology.

4.1.1. Aims

1. Describe *S. flexneri* 2a epidemiology in South Africa
2. Identify endemic *S. flexneri* 2a strain(s) in South Africa
3. Identify imported strains
4. Characterise the AMR profiles of all strains
5. Characterise the virulence gene profiles of all strains
6. Infer how strain accessory genome characteristics influence the epidemiology

4.2. Methods

4.2.1. Selection and sequencing of South African *Shigella flexneri* 2a study isolates

The 286 biochemically identified *S. flexneri* 2a isolates, collected during surveillance of shigellosis in South Africa from 2011 to 2015, inclusive, were initially selected alongside biochemically identified *S. sonnei* isolates (275 isolates), representatively as a complete sample set (total 561 isolates) by year and province (Figure 4.4) [150-154]. Only the *S. flexneri* isolates were then selected for this study.

Biochemical identification was performed as part of the surveillance programme, carried out by The Group for Enteric, Respiratory and Meningeal Diseases Surveillance in South Africa (GERMS-SA), according to laboratory standard operating procedures (SOPs) (Supplementary) [150-154].

Surveillance methods are described in the GERMS-SA annual reports, but in brief involved a collection of public hospitals and the microbiology testing labs which served them, from across all nine provinces (Figure 4.4). With a standard case definition for shigellosis and all included hospitals being required to send in samples for all shigellosis cases for biochemical identification and AMR phenotyping according to laboratory Standard Operating Procedures [150-154].

All study isolates were subjected to whole genome sequencing and quality control as laid out in Sections 2.1 and 2.2 of the Methods chapter (Chapter 2) respectively. Sequence data was received for 273 isolates. The sequence quality of some of the isolates (n=119) was poor so they were re-sequenced at the Centre for Genomic Research (CGR, University of Liverpool) using the Illumina NovaSeq 6000 platform; the DNA library was constructed using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina [37]. Original and new sequence reads were combined for all re-sequenced isolates and then quality trimmed and assessed again according to the standard protocol as laid out in sections 2.1 and 2.2 of the Methods chapter.

Seven study isolates (six re-sequenced) were excluded due to poor sequence quality, two had uneven per base sequence content while five were likely contaminated, indicated by their uneven GC content

curve (Supplementary). Two further isolates were excluded due to poor mapping against a complete *S. flexneri* 2a reference genome (301 strain, chromosome and plasmid: GenBank accessions NC_004337.2 and NC_004851.1), having <20 mean read depth. Four isolates (two re-sequenced) identified, *in silico*, as not *S. flexneri* 2a were also excluded (Supplementary). *In silico* typing was performed with shigaTyper (v1.0.6) and maximum likelihood phylogenetics (Section 2.4 of the Methods chapter). Of the selected isolates, 260 *S. flexneri* 2a were included in this study, 111 were re-sequenced isolates. All unpaired reads were also excluded due to poor per base sequence content.

4.2.2. Sequencing of contextual isolates in this study

A selection of reference isolates from across the known *S. flexneri* global phylogeny were selected for inclusion in the maximum-likelihood phylogeny to provide a global context for the study isolates. The selection criteria and accession numbers for these isolates can be found in the Methods chapter (Chapter 2) Section 2.1.4.

4.2.3. Data collection

Recorded case numbers, taken from the GERMS-SA annual reports, were used to examine the background epidemiology of *S. flexneri* 2a caused shigellosis in the country, the larger sample set from which our study samples were sub-sampled, and the representation of that sample set by our study isolates [150-154]. The annual report from 2014 did not include a breakdown of shigellosis cases and the 2015 report was a reduced report on shigellosis. Due to this missing data, comparisons were only able to be made from 2011 to 2013.

Each province is made up of district municipalities and metropolitan municipalities (both referred to as districts throughout) (Figure 4.4). The degree of urbanisation of these districts was defined according to the methods laid out by the European Commission in March 2020 (https://ec.europa.eu/eurostat/cros/system/files/bg-item3j-recommendation-e.pdf). For this study, provinces determined to be densely populated areas are referred to as ‘urban’, intermediate density

provinces are referred to as 'mixed' and thinly populated provinces as 'rural' (Thomas, personal communication, see acknowledgements above).

Patient age and gender, used to examine the association between age of infection and gender, were collected by hospitals at the time of sample collection according to standard hospital procedures. Some isolates were from blood samples, or from patients who had a *Shigella* positive blood test, these cases were defined as having an invasive or systemic disease presentation.

4.2.4. Global and South African *Shigella flexneri* population structure

The population structure of the study isolates within the known global context was determined using maximum likelihood phylogenetics, as laid out in Section 2.4.1 of the Methods chapter, using reference isolates from across the known *S. flexneri* global phylogeny (Section 2.1.4, Table 4.1) in addition to the South African *S. flexneri* 2a study isolates.

All phylogenies were generated from core-SNP alignments generated with a multiple software pipeline (Section 2.4.1) which mapped quality trimmed sequence reads to a complete reference genome (*S. flexneri* strain 301 chromosome and plasmid: GenBank accessions NC_004337.2 and NC_004851.1), called variant sites against this same reference genome, and then defined a consensus sequence for each isolate. Multiple consensus sequence alignments were run through Gubbins (v2.3.4) to identify core-SNPs. A core-SNP alignment was extracted from the Gubbins output using SNP-sites (v2.4.1).

Two maximum likelihood phylogenies were created, the first was generated from a core-SNP alignment including the complete global *S. flexneri* population reference set (34768 SNPs) and was used to identify which Phylogroup the study isolates belonged to. As all the study isolates were found to belong to Phylogroup 3, a second Phylogroup 3 specific tree was generated to examine population structure and included only Phylogroup 3 reference *S. flexneri* isolates (13279 SNPs) (Table 4.1).

Table 4.1. Shigella flexneri Phylogroup 3 maximum likelihood phylogenetic reference isolates.

Accession	Serotype	Accession	Serotype	Accession	Serotype
ERS025897	Xv	ERS033379	2a	ERS088044	2a
ERS025915	2a	ERS033382	Y	ERS088045	2a
ERS025916	2a	ERS033383	Y	ERS088063	2a
ERS025917	Y	ERS033371	1a	ERS088075	Y
ERS025900	2a	ERS033389	2b	ERS157637	2a
ERS025902	2a	ERS033390	2a	ERS157646	2a
ERS025903	2a	ERS033391	2b	ERS157647	2a
ERS025904	2a	ERS033372	X	ERS157648	2a
ERS025905	2a	ERS033373	2a	ERS157649	2a
ERS025906	2a	ERS033374	2a	ERS157650	2a
ERS025907	2b	ERS033375	2a	ERS157651	2a
ERS025931	2a	ERS033376	2a	ERS157652	2b
ERS025922	2a	ERS093670	5a	ERS157653	2a
ERS025940	X	ERS093673	2a	ERS157654	2a
ERS025941	Yv	ERS093674	2a	ERS157655	2a
ERS025942	1a	ERS093675	2a	ERS157638	2a
ERS025925	2a	ERS093679	2a	ERS157656	2a
ERS025926	2a	ERS093680	2a	ERS157657	2a
ERS025943	2a	ERS093684	2a	ERS157658	2a
ERS025955	2a	ERS093685	2a	ERS157659	2a
ERS025961	2a	ERS093691	Xv	ERS157660	2b
ERS025962	Xv	ERS093698	Y	ERS157661	2b
ERS025964	Yv	ERS093702	Y	ERS157662	2b
ERS025946	Xv	ERS093704	2a	ERS157663	2b
ERS025951	2a	ERS093706	2a	ERS157664	2b
ERS033335	2b	ERS093712	2a	ERS157665	2b
ERS033336	2a	ERS087988	2a	ERS157639	2a
ERS033344	2a	ERS087989	2a	ERS157666	2b
ERS033327	Y	ERS088014	2a	ERS157667	2b
ERS033347	2b	ERS088015	2a	ERS157668	2b
ERS033331	2a	ERS088016	2a	ERS157669	2b
ERS033358	2a	ERS088017	2a	ERS157640	2a
ERS033362	Y	ERS088018	2a	ERS157641	2a
ERS033365	2b	ERS087985	2a	ERS157642	2a
ERS033353	2a	ERS087986	2a	ERS157643	2a
ERS033354	2a	ERS088040	2a	ERS157644	2a
ERS033355	2a	ERS088041	2a	ERS157759	2b
ERS033356	2a	ERS088042	2a	ERS157788	2a
ERS033370	2a	ERS088043	2a		

The South African *S. flexneri* 2a study population structure was also modelled in the context of time, using the second-generation Bayesian Evolutionary Analysis by Sampling Trees software (BEAST2) (v2.6.3). The time tree was generated from a study isolate only core-SNP alignment (7936 SNPs), created in the same way as for the maximum likelihood phylogeny only without the population structure reference isolates. The fasta formatted sequence alignment was converted to nex format with seqmagick (v0.8.0) (https://github.com/fhcrc/seqmagick/).

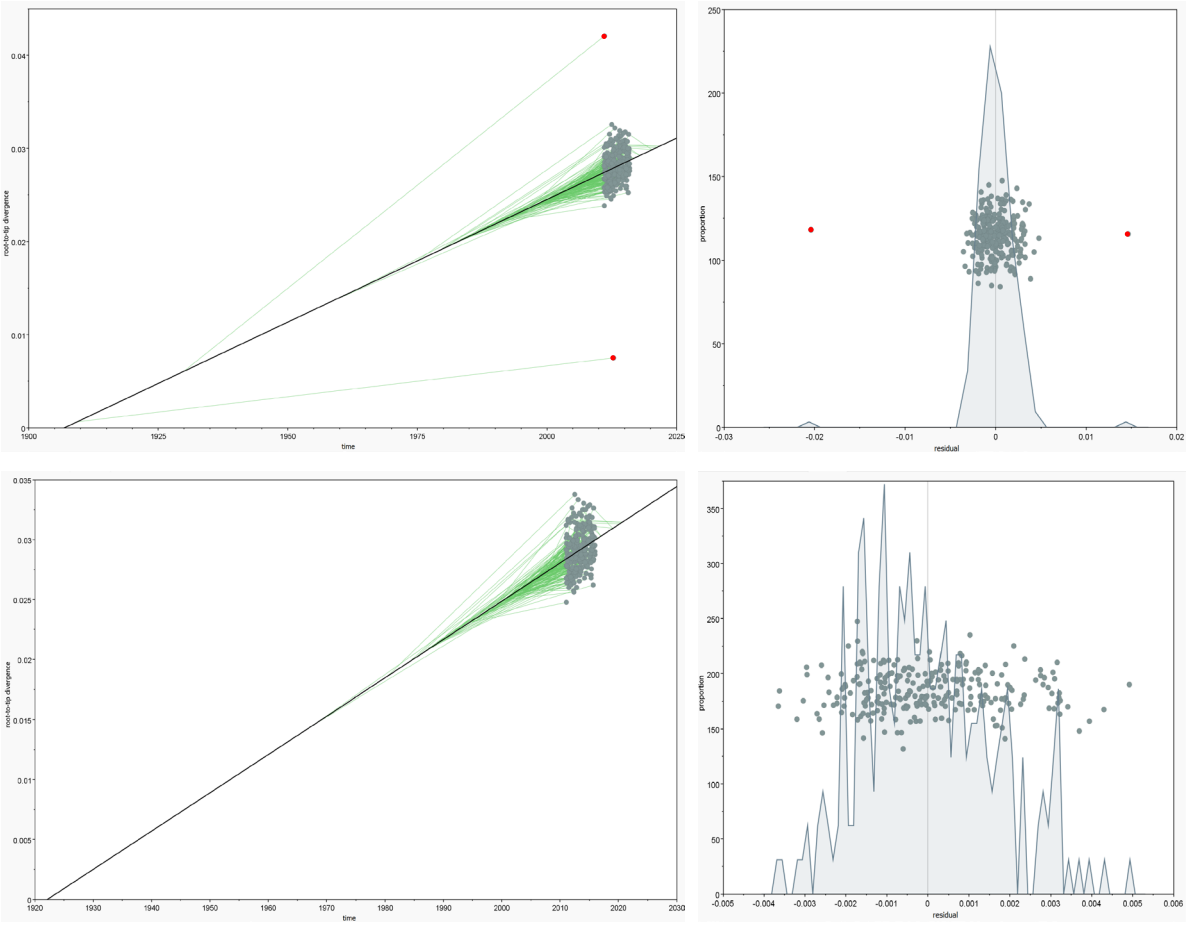

Figure 4.1. Correlation between root-to-tip divergence and of S. flexneri isolates sampling date (left) and the correlation residuals of these isolates (right) – before (top) and after (bottom) outlier isolates were removed. Excluded isolates are marked in red, and green lines show ancestor traces of excluded isolates (top left) and of all isolates (bottom left).

Two isolates (FD01872896, FD01876672) were excluded from the BEAST phylogeny due to being outliers for the molecular clock signal, assessed using the size of the isolate residual in the correlation between phylogeny root-to-tip divergence and isolate sampling date (Figure 4.1). Outliers were

visually identified based on having a large residual distance from the correlation line of best fit as well as having log ancestor traces suggestive of a divergence time widely different from the other isolates (Figure 4.1). The correlation between root-to-tip divergence and sampling date was assessed in TempEST (v.1.5.3) using a RAxML-ng (bootstraps, GTR+G model) generated, *S. flexneri* study isolate-only maximum likelihood phylogeny (according to: http://beast.community/tempest_tutorial) [165].

Being an outlier likely indicates mislabelling of sample dates, contamination, sequence degradation or errors, or recombination. Inclusion of these isolates would obscure the molecular clock signal of the population. Once outlier isolates were excluded, the *S. flexneri* clock rate was in TempEST to be $3.1921e^{-4}$ substitutions per site per year, correlation co-efficient = 0.27.

For the BEAST2 modelling, I used the extended coalescent Bayesian skyline tree model with a relaxed clock model and a log normal prior distribution (prior rate = $1e^{-6}$ as had a single partition, based on https://beast2.blogs.auckland.ac.nz/tag/clock-rates), and site model averaging with BmodelTest, transition-transversion split (prior mutation rate = 1.0, based on: https://beast2.blogs.auckland.ac.nz/tag/clock-rates), to estimate the most appropriate site model from the data [166]. All other priors were left as default [167].

A relaxed clock model allows for variation in evolutionary rate between tree branches instead of assuming the evolutionary rate is constant across the entire tree. Log normal prior distributions vary from normal to negatively skewed, promoting slower evolutionary rates over fast ones while estimating the most appropriate rate distribution from the data.

I ran one MCMC chain with a length of 5,043,000,000 (sampling every 1,000,000), 39% burn-in was removed from the start of every output file to ensure convergence as a spike in the trace was observed. The tree topology was generated from the output trees file with treeannotator (v2.6.3) and then visualised in FigTree (v1.4.4) (https://github.com/rambaut/figtree) and the interactive tree of life (ITOL) online platform and, the former was used to obtain node age estimates and 95% HPDs [158]. A

minimum effective sample size (ESS) of 200 was achieved for all parameters once all runs were completed, combined and burn-in removed, checked using Tracer (v2.6.3) [168].

Population clusters of the South African isolates were defined using RhierBAPS (v1.1.3) the South African *S. flexneri*-only core-SNP alignment in Rstudio (v1.4.1717; R v4.1.0) [169, 170]. The output clustering from this is later referred to as BAPS (Bayesian Analysis of Population Structure) clusters.

4.2.5. Genome assembly

Draft genomes were assembled using Unicycler (v0.4.7) and quality assessed with Quast (v5.0.2), as laid out in Section 3.4 of the Methods chapter [140, 149]. No isolates were excluded on the basis of poor genome assembly.

4.2.6. Antimicrobial resistance profiling

Methods for antimicrobial resistance genotype profiling and phenotype prediction are laid out in Section 2.6 of the Methods chapter. Phenotype data, for resistance against ampicillin, chloramphenicol, streptomycin, tetracycline, cotrimoxazole, nalidixic acid, ciprofloxacin, or ceftriaxone, was used where available. Antimicrobial resistance phenotyping of a proportion of the isolates (Table 2.5) was carried out according to laboratory SOPs (Supplementary).

Where no phenotype data was available it was predicted based on *in silico* detected genotype. Genotyping was carried out using two pieces of software StarAMR (v0.5.1) and AMRfinderPlus (v3.2.3). Phenotyping data was used to optimise the accuracy of the phenotype prediction for this dataset, the methods of optimisation are detailed in Section 2.6.1.

4.2.7. Virulence profiling

A virulence profile was generated for all study isolates by comparing the draft genome assemblies against the VirulenceFinder database, using VirulenceFinder (v2.0.4-1), and against a local, curated virulence genome database (Table 4.2), with BLASTn (v2.10.0+) [155]. The local database of genes was

based on those identified in a previous study [33]. Only identified genes with a ≥99% sequence identity and database gene coverage were accepted as present and functional.

The presence of a selection of virulence loci (SRL-PAI, SRL and pINV) were assessed by assessing the level of read mapping across the region of interest within a selection of reference isolates known to contain the relevant loci (Table 4.3). Quality-trimmed isolate sequence reads were mapped to a reference genome, visually confirmed as containing the virulence loci of interest, with bwa mem (v0.7.17).

Samtools (v1.9) was used to remove unmapped, not primarily aligned, QC failed, duplicate, and supplementary reads, and to sort and index the mapped reads prior to mapping assessment [142, 144]. Confirmation of the virulence loci presence in the respective reference genomes (Table 4.3) was confirmed in ACT following a BLASTn (v2.10.0+) comparison of the mapping reference genome against a reference loci sequence (Table 4.3).

Read mapping to the entire reference genome was also assessed, any isolate with mean read depth ≤10 across the whole reference was excluded from the region of interest mapping analysis, consequently one isolate was excluded from the pINV analysis and six were excluded from the SRL-PAI/SRL analysis.

The read mapping of a positive and a negative control were also assessed (Table 4.3) to ensure that any observed read mapping was due to the presence of the region of interest and not an *in silico* read mapping artefact. Random reads (paired, 100-70 bases, insert size 150-300) were generated from the positive and negative controls (Table 4.3) with the bbmap randomreads.sh script (v38.00) to 60x coverage (www.sourceforge.net/projects/bbmap/). The generated reads were then converted into a set of forward and a set of reverse reads with seqkit (v0.10.1) [171].

Table 4.2. Virulence genes included in curated database.

Gene	Accession number(s)
entA	NC_004337.2 (531914-532660), NZ_CP055292.1 (c618316-617570), NZ_MSJW02000146.1 (c17068-16322), NZ_LPTR01000087.1 (c7619-6873)
entB	NC_004337.2 (531057-531914), NZ_CP055292.1 (c619173-618316), NZ_LPTR01000087.1 (c8476-7619)
entC	NC_004337.2 (528248-529423), NZ_CP055292.1 (c621982-620807), NZ_MSJW02000146.1 (c20554-19379), NZ_LPTR01000087.1 (c11285-10110)
entD	NC_004337.2 (c512813-512184), NZ_CP055292.1 (636788-637408), NZ_LPTR01000126.1 (c1598-978)
entE	NC_004337.2 (529433-531043), NZ_CP055292.1 (c620797-619187), NZ_MSJW02000146.1 (c19369-17750), NZ_LPTR01000087.1 (c10100-8490)
entF	NC_004337.2 (516881-520726), NZ_CP055292.1 (c632710-628829), NZ_LPTR01000126.1 (5676-9557)
fecA	NZ_MSJW02000150.1 (2888-5212)
fecB	NZ_MSJW02000150.1 (5257-6159)
fecC	NZ_MSJW02000150.1 (6156-7154)
fecD	NZ_MSJW02000150.1 (7151-8107)
fecE	NZ_MSJW02000150.1 (8108-8875)
fecR	NZ_MSJW02000150.1 (1848-2801)
fepA	NC_004337.2 (c515219-512979), NZ_CP055292.1 (634373-636613), NZ_LPTR01000126.1 (c4013-1773)
fepB	NC_004337.2 (c527956-527000), NZ_CP055292.1 (622357-623313), NZ_MSJW02000146.1 (20928-21884), NZ_LPTR01000087.1 (11577-12533)
fepC	NZ_CP055292.1 (626668-627483), NZ_LPTR01000087.1 (15870-16685)
fepD	NC_004337.2 (c525665-524649), NZ_CP055292.1 (624678-625682), NZ_LPTR01000087.1 (13880-14884)
fepG	NC_004337.2 (c524652-523660), NZ_CP055292.1 (625679-626671), NZ_LPTR01000087.1 (14881-15873)
fimA	NC_004337.2 (c4382103-4381555), NZ_CP055292.1 (c1363528-1362980), NZ_CP055292.1 (4333156-4333719), NZ_CP055292.1 (c679609-679067), NZ_MSJW02000103.1 (c14019-13477), NZ_LPTR01000053.1 (7046-7594)
fimB	NC_004337.2 (c4384260-4383775), NZ_CP055292.1 (c1365685-1365083)
fimC	NC_004337.2 (c4380914-4380189), NZ_CP055292.1 (c1362339-1361614), NZ_CP055292.1 (c678847-678155), NZ_MSJW02000103.1 (c13257-12565), NZ_LPTR01000053.1 (8235-8960)
fimD	NC_008258.1 (c4350014-4347378), NZ_CP055292.1 (c1361547-1358911), NZ_LPTR01000053.1 (9027-11576)
fimE	NC_004337.2 (c4383180-4382584), NZ_CP055292.1 (c1364605-1364009), NZ_LPTR01000053.1 (5969-6565)
fimF	NZ_CP055292.1 (c1358901-1358371), NZ_LPTR01000053.1 (11586-12116)
fimG	NC_004337.2 (c4376157-4375654), NZ_CP055292.1 (c1358358-1357855), NZ_LPTR01000053.1 (12129-12632)
fimH	NC_004337.2 (c4375634-4374732), NZ_CP055292.1 (c675502-674495), NZ_CP055292.1 (c1357835-1356933), NZ_MSJW02000103.1 (c9912-8905), NZ_LPTR01000053.1 (12652-13554), NZ_LPTR01000196.1 (c1671-664)
iucA	NC_004337.2 (3820792-3822573)
iucB	NC_004337.2 (3822574-3823521)
iucC	NC_004337.2 (3823521-3825263)
iucD	NC_004337.2 (3825260-3826597)
iutA	NC_004337.2 (3826603-3828798)
pic	NC_004337.2 (c3071855-3067737)
sat	NZ_MSJW02000174.1 (772-4659), NZ_LPTR01000224.1 (c1528-95)
set1A	NC_004337.2 (3069744-3070277)
set1B	NC_004337.2 (3069555-3069740)
shiA	NC_004337.2 (3808392-3809435)
shiB	NC_004337.2 (3809950-3810411)
shiC	NC_004337.2 (3811744-3812262)
shiD	NC_004337.2 (3817102-3817500)
shiE	NC_004337.2 (3818599-3819570)
sigA	NC_004337.2 (3060437-3064294)
sitA	NC_004337.2 (c1408785-1407895), NZ_MSJW02000113.1 (20693-21607), NZ_LPTR01000169.1 (c5068-4154)
sitB	NC_004337.2 (c1407895-1407068), NZ_MSJW02000113.1 (21607-22434), NZ_LPTR01000169.1 (c4154-3327)
sitC	NC_004337.2 (c1407071-1406214), NZ_MSJW02000113.1 (22431-23288), NZ_LPTR01000169.1 (c3330-2473)
sitD	NC_004337.2 (c1406217-1405360), NZ_MSJW02000113.1 (23285-24142), NZ_LPTR01000169.1 (c2476-1619)
stx1A	NC_028685.1 (25526-26473), NC_025434.1 (20901-21848)
stx1B	NC_028685.1 (26483-26752), NC_025434.1 (21858-22127), NC_029120.1 (21872-22141)

Table 4.3. Reference sequences and positive and negative controls for assessing the presence or absence of virulence loci.
Bolded = in silico created.

Virulence loci	Reference loci sequence	Mapping reference genome	Positive control	Negative control
pINV	*S. flexneri* 5a M90T strain, pWR501 plasmid (NC_002698.1)	*S. flexneri* 2a 301 strain, chromosome and plasmid (NC_004337.2, NC_004851.1)	*S. flexneri* 2a 301 strain, chromosome and plasmid (NC_004337.2, NC_004851.1)	*S. flexneri* 2a 301 strain, chromosome only (NC_004337.2)
SRL-PAI and SRL	*S. flexneri* 2a YSH6000 strain, SRL-PAI (AF326777.3)	*S. flexneri* 89-141 strain, chromosome (CP026803.1) and plasmid (CP026804.1)	***S. flexneri* 2a 301 strain, chromosome and plasmid (NC_004337.2, NC_004851.1) + *S. flexneri* 2a YSH6000 strain, SRL-PAI (AF326777.3)**	*S. flexneri* 2a 301 strain, chromosome and plasmid (NC_004337.2, NC_004851.1)

Table 4.4. Coordinates of 'region of interest' used in the virulence loci read mapping analysis.

Virulence loci	Coordinates of virulence locus in mapping reference
pINV	NC_004851.1, entire plasmid sequence
SRL-PAI	CP026803.1, positions: 2458488-2525183
SRL	CP026803.1, positions: 2473016-2488021

The presence, or absence, of the virulence loci was assessed by determining 1) mean read depth across the region of interest (Table 4.4), and 2) breadth of coverage across the region of interest, defined as the number of bases with a read depth >1 divided by the total number of bases. Complete absence was defined by a mean read depth of 0 or the breadth of mapping coverage ≤5% of the region of interest. Presence, full or partial, was categorised based on the breadth of mapping coverage.

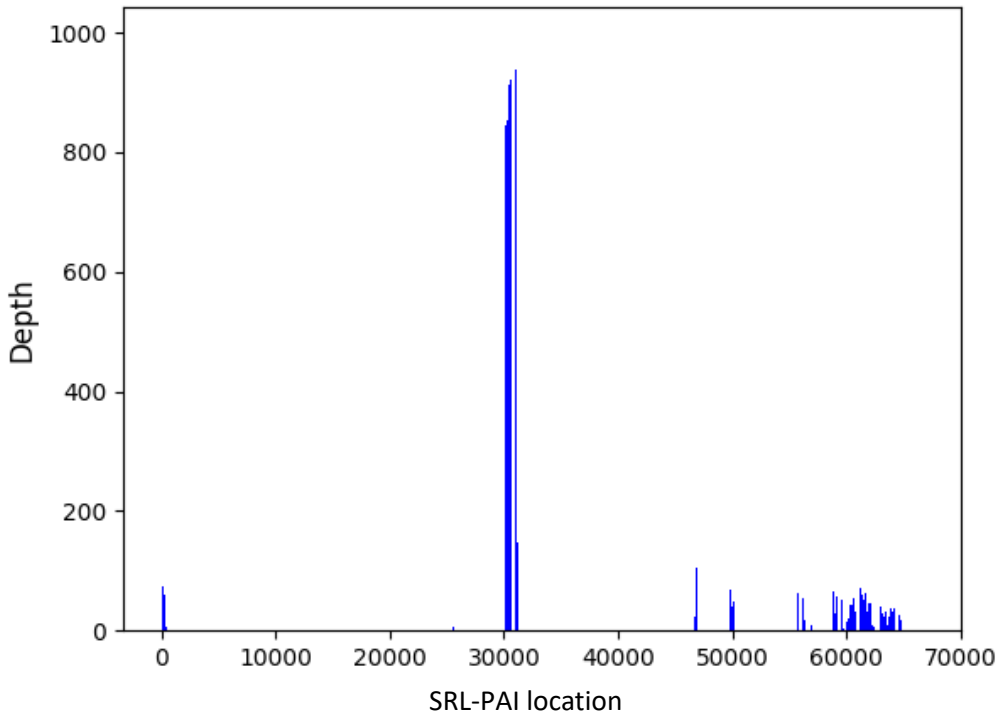

Figure 4.2. Negative control read mapping across the SRL-PAI.
Mean read depth = 19 and breadth of coverage = 14%.

The exception to this was the SRL-PAI as the negative control had a mean read depth of 19 and breadth of 14%. Visual examination of the mapping showed a mapping peak around position 30000 which accounts for the high mean read depth (Figure 4.2). As the peak was not observed in the study isolates but was in the positive control, this mapping pattern may be specific to the reference isolate used to generate the control reads. Regardless the absence threshold was adjusted to coverage ≤5% for the SRL-PAI.

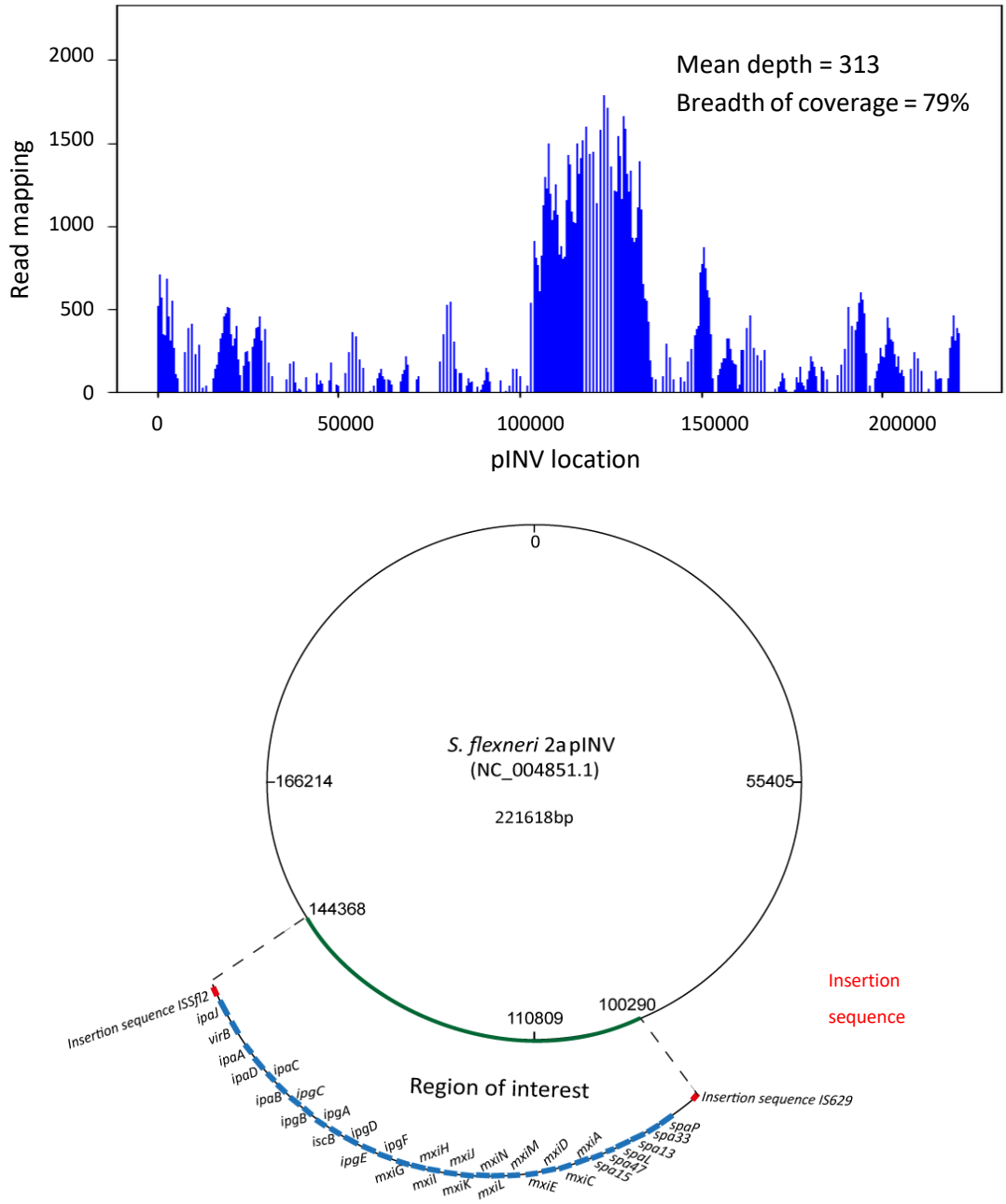

Figure 4.3. Example large virulence plasmid (pINV) read mapping graph (for study isolate FD01872878) (top) and a representation of the pINV 'region of interest' location and encoded genes (bottom).
Graphs (example above, top) of read mapping depth, of study isolates reads, at each base in reference pINV was used to visually identify region of duplication. Region of interest, based on the observed region of duplication, was defined on reference pINV (above, bottom). Region of interest contains virulence-associated mxi-spa locus.

Read depth at each site of the virulence locus of interest (Table 4.4) was determined using bedtools (v2.29.2). The mean read depth, breadth of coverage and a graph of the read depth across the pINV for each isolate were defined using an in-house python script (Supplementary).

The in-house read mapping summary stats python script (Supplementary) also generated read depth mapping graphs. From these read depth graphs, I observed a region of the pINV which had a much greater read depth than the rest of the plasmid in several isolates (Figure 4.3).

To further examine the relative mapping depth across the 'region of interest' compared to the rest of the pINV, I assessed the mean read depth and breadth across this region specifically. This region of interest was defined as GenBank accession: NC_004851.1, Positions: 100290-144368; starting at a mobile genetic element, covering the mxi-spa locus, and ending in a mobile genetic element (Figure 4.3). Read mapping across this region was assessed in the same way as laid out above for the entire pINV.

4.2.8. Population dynamics

Dynamics of the *S. flexneri* 2a population were explored using the same BEAST2 extended coalescent Bayesian skyline model as was used for the population structure (Section 4.2.4).

The Bayesian skyline model estimates the population size between multiple time points across the modelled phylogeny and produces a plot of these estimated population sizes, which together show the changing population size through time [166]. In the standard version of this model, the number of time points and/or the amount of time between them is decided by the user. The extended version used here estimates, from the data and priors, the most appropriate number of time points and the amount of time between them [172].

It is an appropriate model for our data as we have sampled a small proportion of the total population and the coalescent method assumes a small sample size. The model also assumes random mixing of the populations, which is unlikely to be completely met as no population will mix completely randomly. However, the high proportion of sampling from urban regions likely increases the amount of population mixing, as does the use of recently sampled isolates.

An attempt at modelling structured population dynamics, between defined sub-populations (based on age and gender of patients) and geographic regions (rural, mixed, and urban districts), was made using two modelling methods. The first was with marginal approximation of the structured coalescent (MASCOT), a structured coalescent model which approximates rather than infers ancestral migration histories and thus requires less computation [173, 174]. This method resulted in an error message "too many iterations, return negative infinity" which was likely caused by an increasing population size with time when the model assumes a constant population size.

Following the failure of MASCOT, the MultiTypeTree structured coalescent model was used. This model infers rather than estimates ancestral migration, taking longer to run but can model structured population dynamics for a wider range of population dynamics [175]. In this case, the model would not converge, even when a sub-sample of isolates was used to reduce the number of isolates by 30%.

4.2.9. Statistics

Most variable associations were compared with a chi-squared test of association, carried out using the raw numbers but reported, in most cases, as a percentage difference. For the background epidemiology and sample set representativeness, these percentage differences are defined as the observed cases or samples (O) as a fraction of the national total (T) minus the expected (E) fraction of the national total ($\frac{O}{\mathrm{To}} - \frac{E}{\mathrm{Tc}}$), where T_o is the total observations (cases or samples) and T_c is the total reported cases. For associations between isolate or patient metadata associations, the percentage difference is defined as the difference between observed and expected number isolates as a fraction of the expected number of isolates ($\frac{O-E}{E}$).

Association between AMR determinant or profile and geographic region was tested with a Fisher's exact test [176, 177]. Chi-squared tests were used to examine the associations between AMR determinants or profile and *Shigella* sub-population. Meanwhile, for associations between virulence gene and *Shigella* sub-population, odds ratios and two-tailed Z tests were used to assess the

probability of results being due to chance. The Bonferroni correction was used to adjust the statistical significance threshold for multiple comparisons [178].

4.3. Results

4.3.1. Epidemiology and evaluating representativeness of isolates

Analysis of the epidemiological case reporting statistics from the GERMS-SA annual reports revealed higher than expected reported numbers of shigellosis cases relative to population size in some geographical areas [150-152]. Specifically, reported cases were much higher than would be expected in Gauteng (+13.9%) and Western Cape (+11.8%), while the converse was true of Limpopo (-9.8%), Mpumalanga (-5.2%), North West (-5.7%) and KwaZulu-Natal (-5.6%; $\chi^2(26, n=5078.01) = 2156$, $p=0.000$), than if the cases were distributed equally with the total population living in each province (as reported in the GERMS-SA annual report). This might represent an imperfect surveillance system or genuine differences in the incidence of the disease in different geographical areas.

To assess the representativeness of the sample set of reported shigellosis, I compared the number of isolates to the expected number of isolates based on reported case numbers. The number of isolates in the complete sample set (both *S. sonnei* and *S. flexneri* samples) was representative of the reported case numbers (2011-2013) by both province ($\chi^2(26, n=353) = 19.02$, $p=0.890$) and year (2011-2013; $\chi^2(5, n=353) = 2.75$, $p=0.474$) (Figure 4.4A), indicating that the initial sample set selection was proportional to disease burden.

When assessing the representativeness of the *S. flexneri* 2a sub-set, I found the distribution of *S. flexneri* isolates was as expected by year (2011-2013, $\chi^2(2, n=168) = 2.861$, $p=0.239$), but not by province, given the reported cases ($\chi^2;(8, n=260) = 31.091$, $p=0.0001$); the number of isolates in Gauteng was 13% lower than expected but 11% greater in Western Cape (Figure 4.4B). This aligns with the higher proportion of cases attributed to *S. flexneri* 2a in Western Cape, compared to other serotypes and provinces, seen in the GERMS-SA reports [150-154, 162-164]. The observed association between *S. flexneri* province is likely due to serotype-specific geographic distribution within the country.

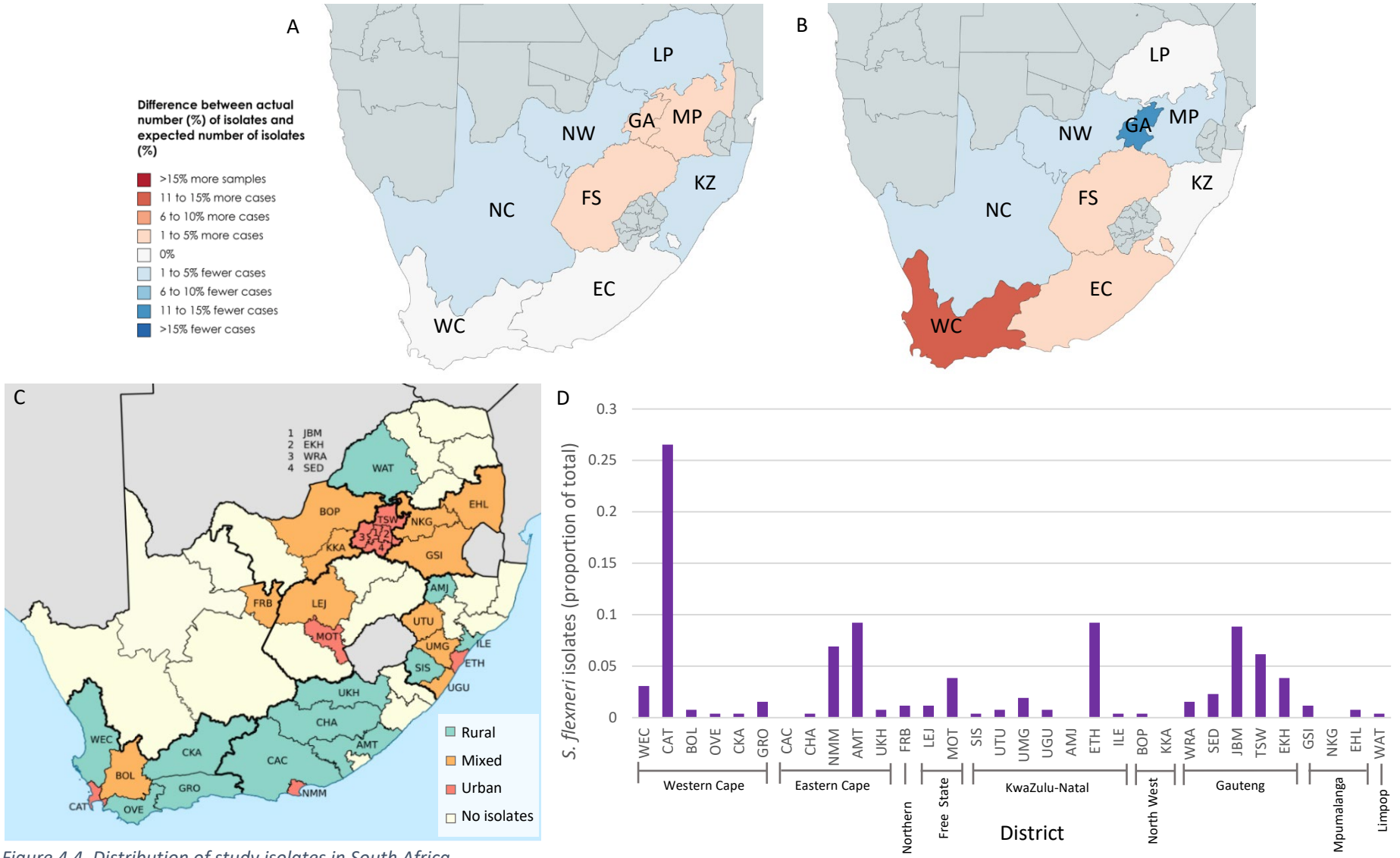

Figure 4.4. Distribution of study isolates in South Africa

A. Difference between the actual and expected number of isolates (*S. flexneri* and *S. sonnei*) by province, based on recorded shigellosis cases from 2011 to 2013. **B.** Difference between the actual and expected number of *S. flexneri* 2a isolates by province, based on recorded shigellosis cases from 2011 to 2013. **C.** Urbanisation level of sampled districts. **D.** Distribution of *S. flexneri* 2a isolates across sampled districts, n = 260. WEC = West Coast DM, CAT = City of Cape Town MM, BOL = Cape Winelands DM, OVE = Overberg DM, CKA = Central Karoo DM, GRO = Garden State DM, CAC = Sarah Baartman DM, CHA = Chris Hani DM, NMM = Nelson Mandela DM, AMT = Amathole DM, UKH = Joe Gqabi DM, FRB = Frances Baard DM, LEJ = Lejweleputswa DM, MOT = Manguang MM, SIS = Harry Gwala DM, UTU = uThukela DM, UMG = , UGU =Ugu DM , AMJ = Amajuba DM, ETH = City of eThekwini MM, ILE = iLembe DM, BOP = Bojanala Platinum DM, KKA = Dr. Kenneth Kaunda MM, WRA = West Rand DM, SED =Sedibeng DM , JBM = City of Johannesberg MM, TSW = City of Tshwane MM, EKH = Ekurhuleni MM, GSI = Gert Sibande DM, NKG = Nkangala DM, EHL = Ehlanzeni DM, WAT = Waterberg DM.

Eleven isolates were isolated from blood samples, indicating an invasive disease presentation. Invasiveness was found to be associated with *S. flexneri,* as opposed to *S. sonnei,* (9 isolates, $\chi^2(1, n=513) = 4.36$, $p=0.037$). The number of *S. flexneri* invasive disease-presenting isolates was too small to assess the potential associations between invasiveness and sample year, age, gender, or province.

4.3.2. Population structure

To contextualise the South African *S. flexneri* 2a isolates, a phylogenetic tree including representatives from across the known global *S. flexneri* phylogeny was constructed (Sections 2.1.4 and 4.2.2). This revealed that all the study isolates were part of global Phylogroup 3. A second Phylogroup 3 specific phylogeny showed that the study isolates were split across two main sub-clusters with some more distantly related isolates (Figure 4.5A). BAPS clustering was consistent with these clusters, the smaller of the two main population sub-clusters was defined as a single BAPS cluster (BAPS1) while the larger of the main sub-clusters was grouped into three BAPS clusters (BAPS 2, 3 and 4) (Figure 4.5). BAPS cluster 3 also included the more distantly related isolates (Figure 4.5).

Phylogenetic dating revealed that the most recent common ancestor (MRCA) of the BAPS2, 3 and 4 sub-population was late-1983 (95% HPD = early 1979 to mid-1988), making it older than BAPS1 which dated to mid-1992 (95% HPD = mid-1989 to late 1995) (Figure 4.5C). The MRCA of all study *S. flexneri* 2a isolates was present in late 1956 (95% HDP = late 1945 to late 1965). Making the MRCA of all the South African isolates younger than the previous estimated MRCA of Phylogroup 3 the median date of which was 1848 [33].

Both BAPS2 and 4 are younger sub-clusters within BAPS3, which likely emerged around the same time; BAPS4 in mid-2001 (95% HPD = early 1999 to late 2003) and BAPS2 in early 1994 (95% HPD = mid-1991 to early 1997) (Figure 4.5C).

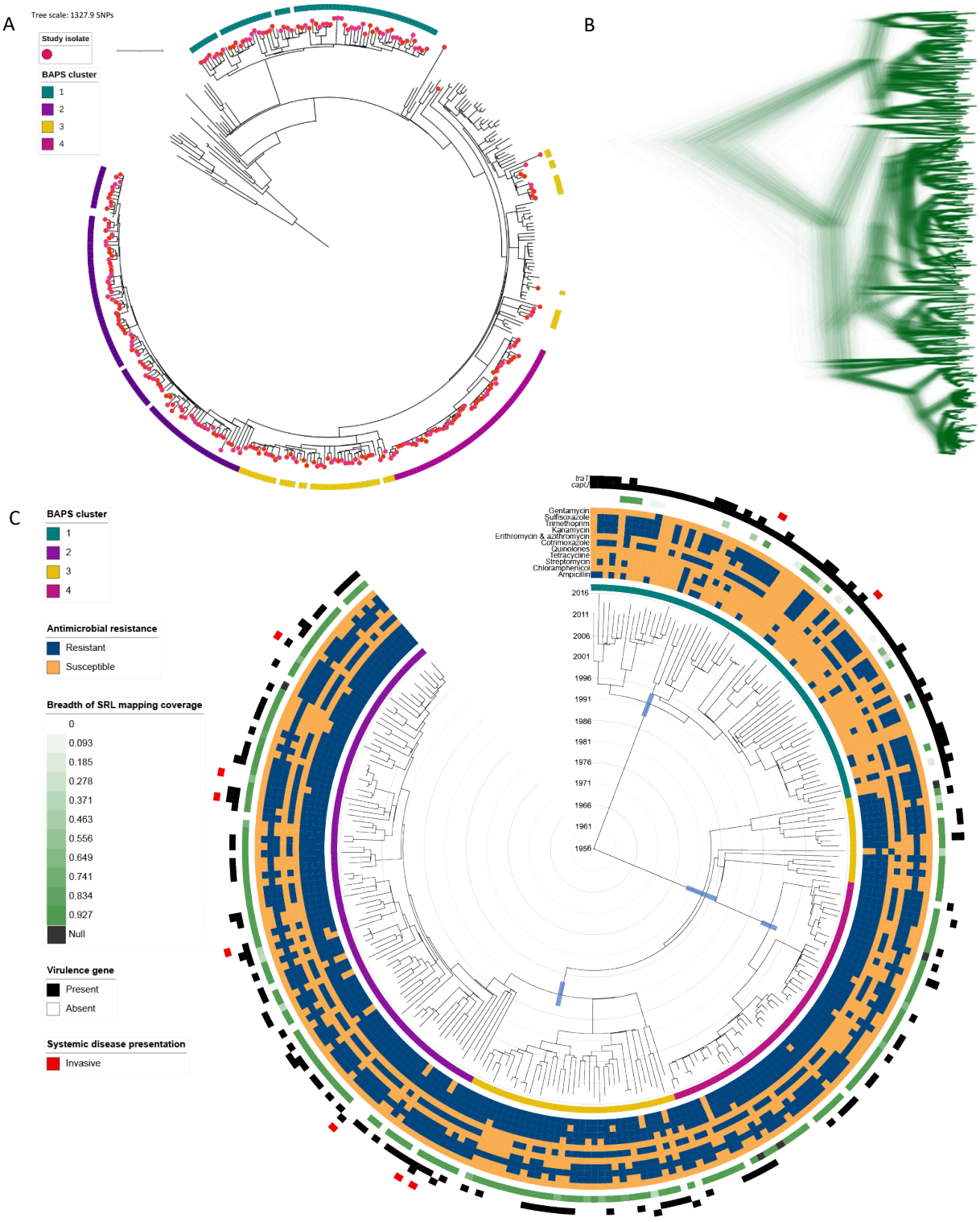

Figure 4.5. Population structure and clustering of predicted antimicrobial resistance phenotypes and virulence genotypes of S. flexneri 2a.

A. Maximum likelihood phylogeny population structure of known global phylogroup 3, study isolates marked with pink terminal nodes and BAPS clusters shown by colour in external ring. **B.** Statistical support for BEAST generated tree topology, shown in C, visualised in densiTree. **C.** BEAST generated phylogeny, with branch length corresponding to time and tree scale behind tree in concentric rings, 95% HPDs on BAPS cluster MRCAs shown as blue bars. BAPS cluster indicated by colour in inner most external ring. Antimicrobial resistance phenotype (predicted/tested) shown in blue (resistant) / peach (susceptible) block of external rings. Read mapping breadth across the reference SRL shown by green external ring. Presence of capU and traT virulence genes indicated by black external ring. Detection of isolate in patient blood shown by red in outer most external ring.

4.3.3. Antimicrobial resistance

Increasing antimicrobial resistance is a hindrance to continued effective treatment of shigellosis. I found that multidrug resistance (MDR), defined as resistance to three or more antimicrobial classes, was widespread (91.5% of *S. flexneri* isolates) (Figure 4.6A).

Some diversity in the identified resistance profiles was observed, with the most common resistance profile (resistance to trimethoprim, streptomycin, ampicillin, chloramphenicol, tetracycline, sulfisoxazole and kanamycin) being shared by only 38% of isolates (Figure 4.6A). However, only four isolates were predicted to be resistant to other antimicrobials (azithromycin, gentamycin, or ciprofloxacin) (Figure 4.6A).

While resistance to quinolones (nalidixic acid) was detected phenotypically (1 isolate), no resistance determinant could be identified. The incongruency between quinolone resistance phenotype and genotype in this isolate may be due to human error in the phenotype recording or novel mechanisms of resistance in the isolate.

For most antimicrobials, a single gene was the dominant resistance determinant present in the population (Figure 4.6B, Table 4.5). Predicted pan-susceptibility was seen in 19 of the 260 isolates. A positive association between drug susceptibility and Free State province was observed, both as a negative association with MDR (55/64 isolates were MDR, 24% lower than expected, fisher's exact two-tailed p=0.008) and as a positive association with pan-susceptibility (3/13 isolates, 216% greater than expected, Fisher's exact two-tailed p=0.043).

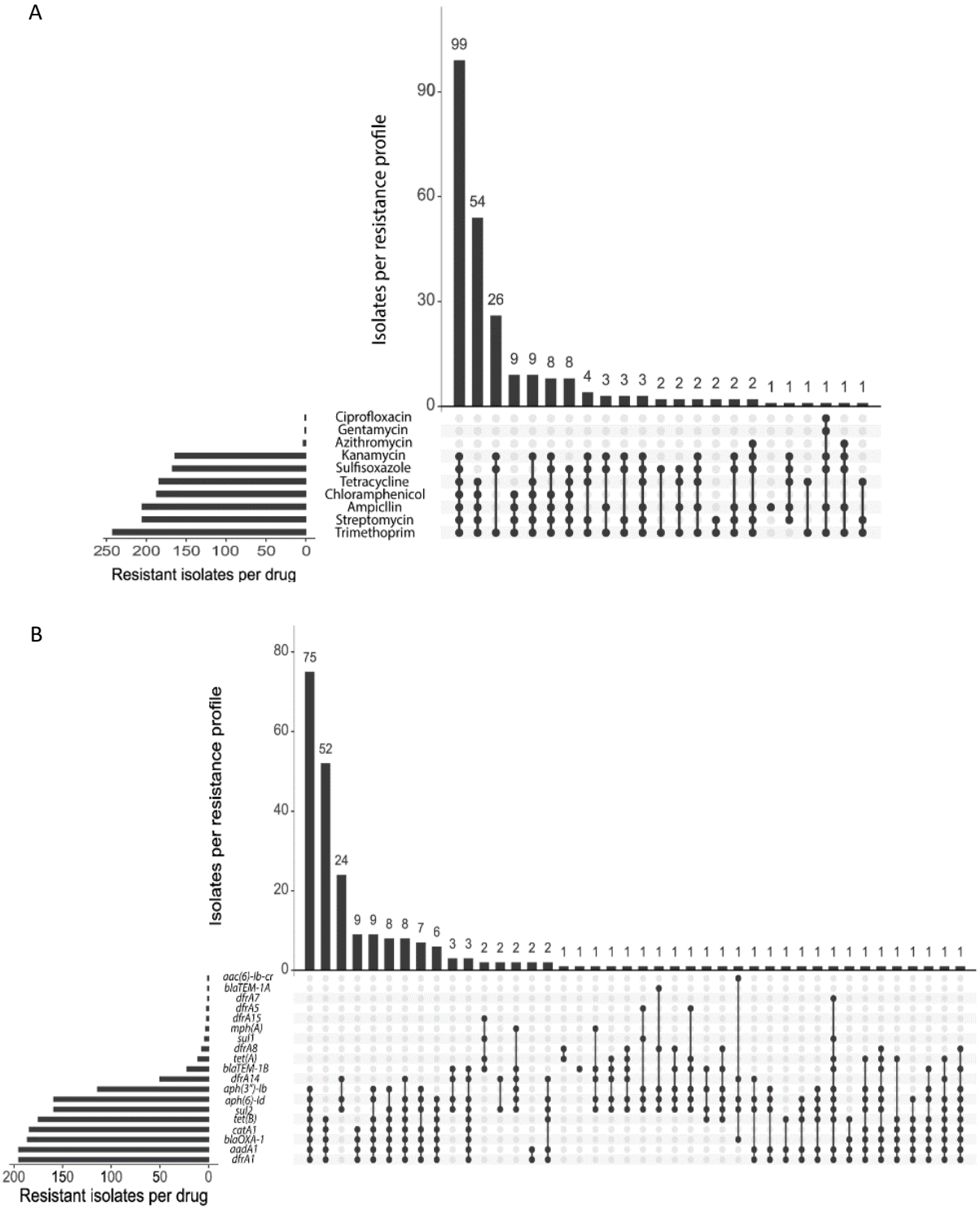

Figure 4.6. Antimicrobial resistance profiles in S. flexneri 2a.

A. Predicted phenotypic profiles. **B.** Predicted genotypic profiles. In both figure parts, combinations of antimicrobials (A) to which resistance was predicted (or tested where data available) or AMR determinants (B) are indicated by black dots to the right of the drug/determinant name, connected by a vertical black line. Vertical histogram along top of each figure part shows the number of isolates with each AMR combination (n above bar). Horizontal histogram to left of each figure part shows the number of isolates resistant to each antimicrobial (A) or with each resistance determinant (B)

Table 4.5. Antimicrobial resistance gene presence in the S. flexneri 2a sample set.
^ only identified by AMRfinder

Genotype / phenotype	**Antimicrobial**	**Resistance determinant**	**Isolates with determinant** n (% of total)
Phenotype & genotype	Ampicillin	*blaTEM-1A* *blaTEM-1B* *blaOXA-1*	1 (0.4) 32 (12.3) 186 (71.5)
Phenotype & genotype	Chloramphenicol	*catA1*	185 (71.2)
Phenotype & genotype	Tetracycline	*tet(A)* *tet(B)*	11 (4.2) 3 (1.2)
Phenotype & genotype	Streptomycin	*aph(3'')-Ib*	114 (43.8)
Genotype	Erythromycin and Azithromycin	*mph(A)*	3 (1.2)
Genotype	Kanamycin	*aph(6)-Id*	159 (61.2)
Genotype	Gentamycin	*aac(6)-Ib-cr*	1 (0.4)
Genotype	Trimethoprim	*dfrA1* *dfrA5* *dfrA7* *dfrA8* *dfrA14* *dfrA15*	195 (75) 2 (0.8) 1 (0.4) 7 (2.7) 50 (19.2) 2 (0.8)
Genotype	Sulfisoxazole	*sul1* *sul2*	4 (1.5) 159 (61.2)
Genotype	Streptothricin^	*sat2*	195 (75.0)
Genotype	Macrolide-lincosamide-streptogramin B^	*ermD*	249 (96.8)

4.3.4. Virulence

To examine virulence in the population, virulence genotype profiles were determined for all isolates. Several known virulence loci were highly prevalent in the population (Table 4.6). The *fim* locus encodes type 1 fimbriae [179-181]. While the enterobactin genes are involved in iron metabolism [182, 183].

Breadth of mapping to the complete SRL-PAI, a known virulence and MDR encoding pathogenicity island, was between 0 and 40%, indicating the complete SRL-PAI was not present in a single *S. flexneri* isolate. The PAI was likely completely absent in 30.7% of isolates (<14% breadth of coverage, threshold based on breadth of mapping observed in negative control).

Table 4.6. Virulence gene prevalence by serotype

Virulence Locus	Virulence gene	Number of isolates (% of total isolates)
SHI-1	*pic*	260 (100.0)
	sigA	260 (100.0)
SHI-2	*shiA*	260 (100.0)
	shiB	248 (95.4)
	shiC	256 (98.5)
	shiD	213 (81.9)
	shiE	260 (100.0)
	iucA	259 (99.6)
	iucB	259 (99.6)
	iucC	259 (99.6)
	iucD	259 (99.6)
	iutA	259 (99.6)
Enterobactin	*entA*	229 (88.1)
	entB	229 (88.1)
	entC	229 (88.1)
	entD	231 (88.8)
	entE	229 (88.1)
	entF	235 (90.4)
	fepA	235 (90.4)
	fepB	229 (88.1)
	fepC	235 (90.4)
	fepD	235 (90.4)
	fepG	235 (90.4)
fim	*fimA*	260 (100.0)
	fimB	260 (100.0)
	fimC	260 (100.0)
	fimD	260 (100.0)
	fimE	260 (100.0)
	fimF	260 (100.0)
	fimG	260 (100.0)
	fimH	260 (100.0)
	capU	154 (59.2)
	celb	3 (1.2)
	gad	77 (29.6)
	ipaD	253 (97.3)
	lpfA	207 (79.6)
	mcbA	1 (0.4)
	ompT	1 (0.4)
	sat	259 (99.6)
	virF	260 (100.0)

Mapping to the SRL alone was also variable, between 0 and 81%, though the coverage was high enough to suggest full or partial SRL presence in 71.5% of isolates (Figure 4.5C). Mapping coverage was 87.8% in the positive control suggesting that less than 100% mapping coverage does not mean only partial presence of the SRL.

The large virulence plasmid (pINV) was detected in all *S. flexneri* isolates (mean read depth ≥9 and breadth of coverage ≥70%). The breadth of mapping in the positive control being only 70% (mean read depth = 39) shows that, again, mapping coverage less than 100% does not mean only partial presence of the pINV.

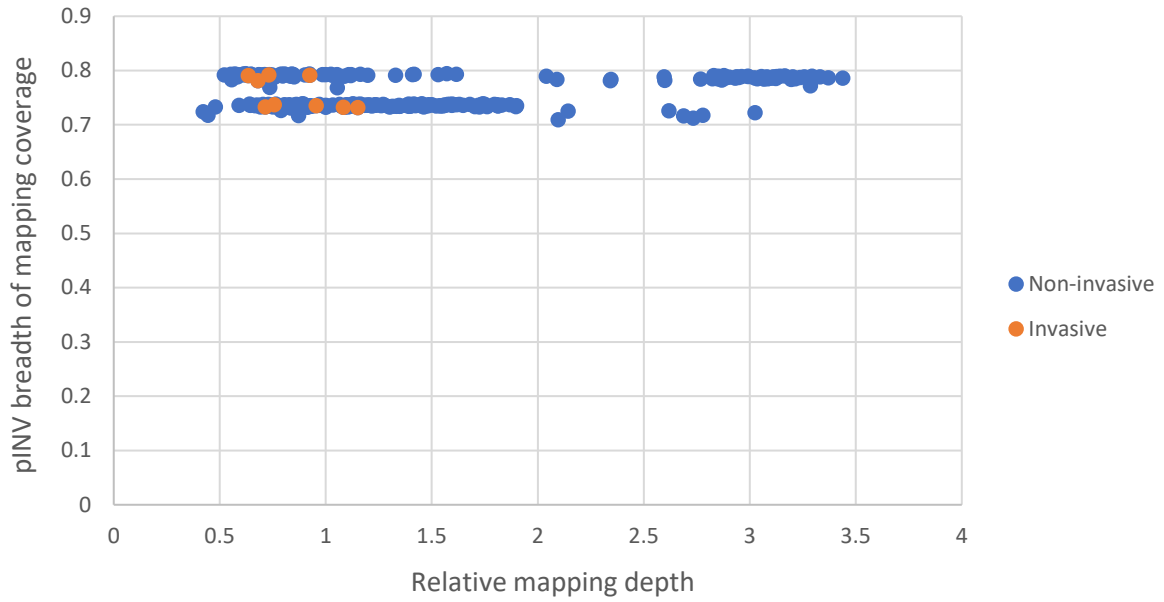

Figure 4.7. The mapping depth across the virulence-associated region of interest in the pINV relative to the mapping depth across the whole pINV, compared to the breadth of mapping coverage across the pINV.
The region of interest has undergone at least one duplication event in a minority of isolates. No association was found between 'region of interest' multiplication and invasive disease presentation.

A 'region of interest' within the pINV with high read depth in some isolates was identified (Figure 4.3). To examine possible duplication of this region, encoding the mxi-spa locus, the mean read depth across this region was compared to the mean read depth across the entire pINV (Figure 4.7). There was evidence of at least one duplication event of all or part of this region in a minority of isolates; 93 isolates (33.8%) had a relative read mapping depth across the 'region of interest' ≥1.5x that of the

complete pINV, 67 (25.9%) had a relative mapping depth ≥1.75, while 55 (21.2%) had a relative mapping depth ≥2.0. No association between region duplication and a systemic disease presentation (invasiveness) was found (Figure 4.7).

4.3.5. Sub-population resistance, virulence and associations with systemic disease

To examine the role of AMR or virulence in the evolution of *S. flexneri* 2a in the country I examined the associations between the identified sub-clusters with identified AMR and virulence determinants. BAPS1 was found to be associated with susceptibility to ampicillin (46 isolates, $\chi^2(1, n=260) = 126.51$, p=0.0000), chloramphenicol (63 isolates, $\chi^2(1, n=260) = 191.70$, p=0.0000), tetracycline (54 isolates, $\chi^2(1, n=260) = 201.85$, p=0.0000), and streptomycin (64 isolates, $\chi^2(1, n=260) = 138.38$, p=0.0000) specifically as well as containing all 19 pan-susceptible isolates (Figure 4.5). While BAPS2, 3 and 4 were associated with MDR (three or more antimicrobial classes), the only isolates not MDR belonged to BAPS1.

Table 4.7. Associations between virulence gene and phylogenetic cluster in S. flexneri 2a.
P-values were calculated with a two-tailed z test. Bolded odds ration have p-value > 0.05

Gene	BAPS cluster	Proportion of cluster with gene	Odds ratio			P-value		
			Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
capU	1	1	1.00					
	2	0.54	**1.84**	1.00		0.0000		
	3	0.37	**1.84**	1.49 (3.00-0.74)	1.00	0.0000	-	
	4	0.38	**2.60**	1.42 (2.84-0.71)	0.95 (2.10-0.43)	0.0000	-	-
traT	1	0.34	1.00					
	2	0.15	**2.21 (4.77-1.02)**	1.00		0.0065		
	3	0.12	**2.98 (8.06-1.10)**	1.35 (3.75-0.48)	1.00	0.0043	-	
	4	0.04	**8.94 (24.17-3.30)**	4.04 (18.57-0.88)	3.00 (15.62-0.58)	0.0001	-	-

The negative association of the SRL (which encodes ampicillin, chloramphenicol, tetracycline and streptomycin resistance) with the susceptible BAPS1, and the positive association of the SRL with BAPS2, 3 and 4, point towards the presence/absence of the SRL being the cause of the drug resistance levels of these sub-populations (Figure 4.5). BAPS1 was, however, associated with virulence genes *capU* and *traT* (Table 4.7 and Figure 4.5).

Invasiveness, or systemic disease presentation, was found to be positively associated with BAPS2 and negatively associated with BAPS3 and 4 which both lacked invasive isolates (Fisher's exact p=0.038) (Figure 4.5C).

Genome recombination, detected with Gubbins, was identified between bases 3846033-3849962 of *S. flexneri* 301 strain reference genome in all but BAPS1. This region, containing *yieC*, *yieL* and *yieK* genes, encodes three outer membrane proteins which are probably involved in carbohydrate transport and metabolism, based on protein structure as recorded in the UniProt database [184]. While recombination between bases 3861515-3863531 was detected solely in BAPS1, the encoding region of tryptanophase *tnaA*.

4.3.6. Sub-population geographic distribution

To determine if differences in geographical distribution exist between the identified sub-populations, I examined the association between province and BAPS cluster. An overall association between BAPS cluster and province was identified (Fisher's exact p=0.0005) (Figure 4.8).

Only BAPS1 and 4 were individually found to have a geographical distribution significantly different from expected based on the reported case distribution (Fisher's exact p=0.0857 and p=0.0257, respectively). When examining BAPS cluster individually the sample sizes from some provinces are small and it may be that there is not the power to detect geographical distribution difference for the other BAPS clusters.

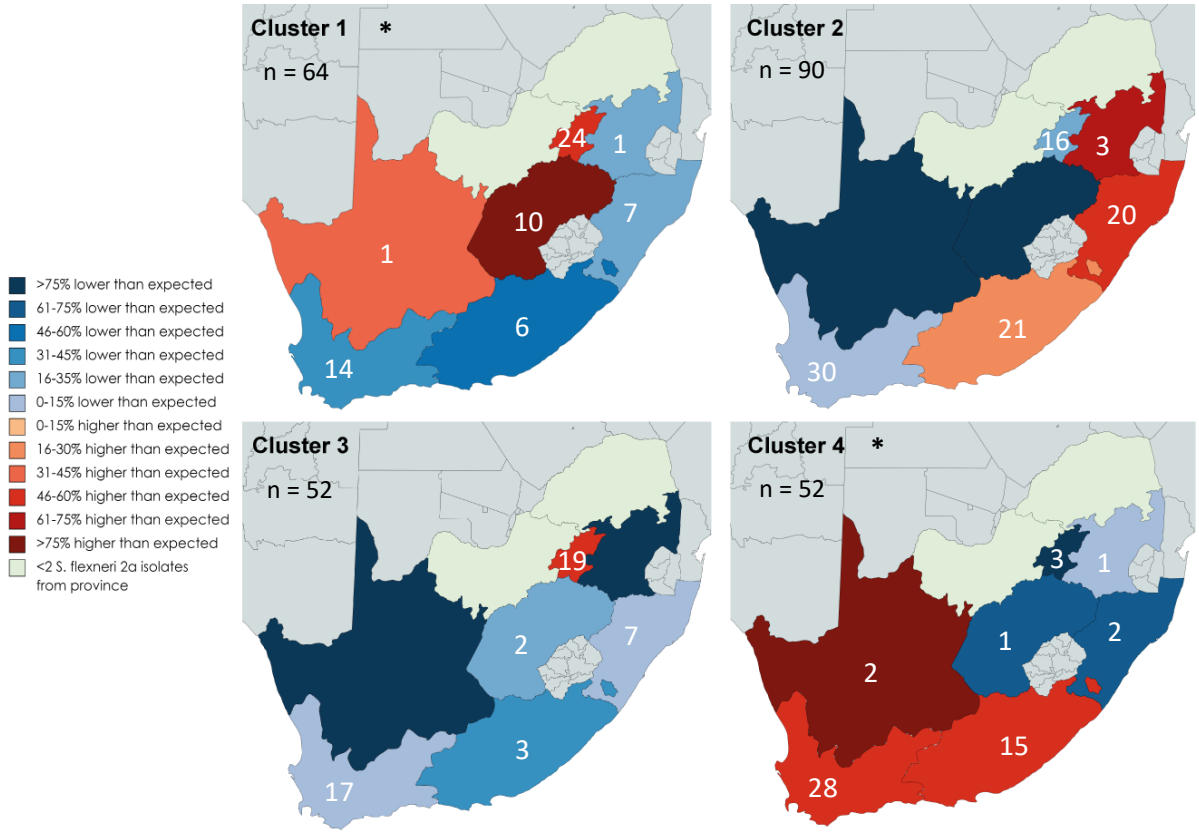

Figure 4.8. S. flexneri sub-population geographic associations.

The number of isolates of each BAPS cluster from each province displayed as a percentage of the expected number. The expected number of isolates is based on the proportion of total isolates from province multiplied by the total number of isolates in the relevant BAPS cluster. Two provinces had fewer than two *S. flexneri* 2a isolates present in the sample set (across all BAPS clusters) and were excluded from the analysis. * Fischer's exact test found this cluster to have a distribution significantly different from expected P<0.05.

Table 4.8. S. flexneri population clusters by degree of district urbanisation and comparison of observed with expected.

Where expected numbers are calculated as the proportion of cluster isolates from total isolates multiplied by the number of isolates at each level of urbanisation. Red fill = greater than expected number of isolates, blue fill = lower than expected.

BAPS cluster	***Isolates***			***Observed difference to expected (% of expected)***		
	Rural	***Mixed***	***Urban***	***Rural***	***Mixed***	***Urban***
1	4	9	48	-65%	73%	8%
2	25	5	59	50%	-34%	-9%
3	4	1	43	-55%	-76%	23%
4	13	6	29	45%	46%	-17%

The uneven geographic distribution of BAPS1 is further supported by the negative association between Free State province and MDR (Fisher's exact p=0.0228); BAPS1 being negatively associated with MDR but positively associated with Free State province (Figure 4.8).

An association between BAPS cluster and level of district urbanisation was observed (Fisher's exact p=0.0001). Specifically, BAPS1 was negatively associated with rural districts but positively associated with mixed districts, the converse of BAPS2 (Table 4.8). Likewise, BAPS3 was negatively associated with rural and mixed districts, the converse of BAPS4 which was positively associated with both rural and mixed districts (Table 4.8).

4.3.7. Population dynamics

To examine how the population size changed through time a coalescent skyline plot was created (Figure 4.9A). It showed that the population size likely increased from 2001 to 2009. The uncertainty during the last few years of the study period means that we cannot be certain of changes in population size after around 2009. While further into the past, beyond 2005, the number of tree events decreases (Figure 4.9B). Fewer tree events means that the relative influence of the data decreases compared to that of the priors, the accuracy of the model, therefore, likely also decreases; beyond 2001 we should be sceptical of the results of the model.

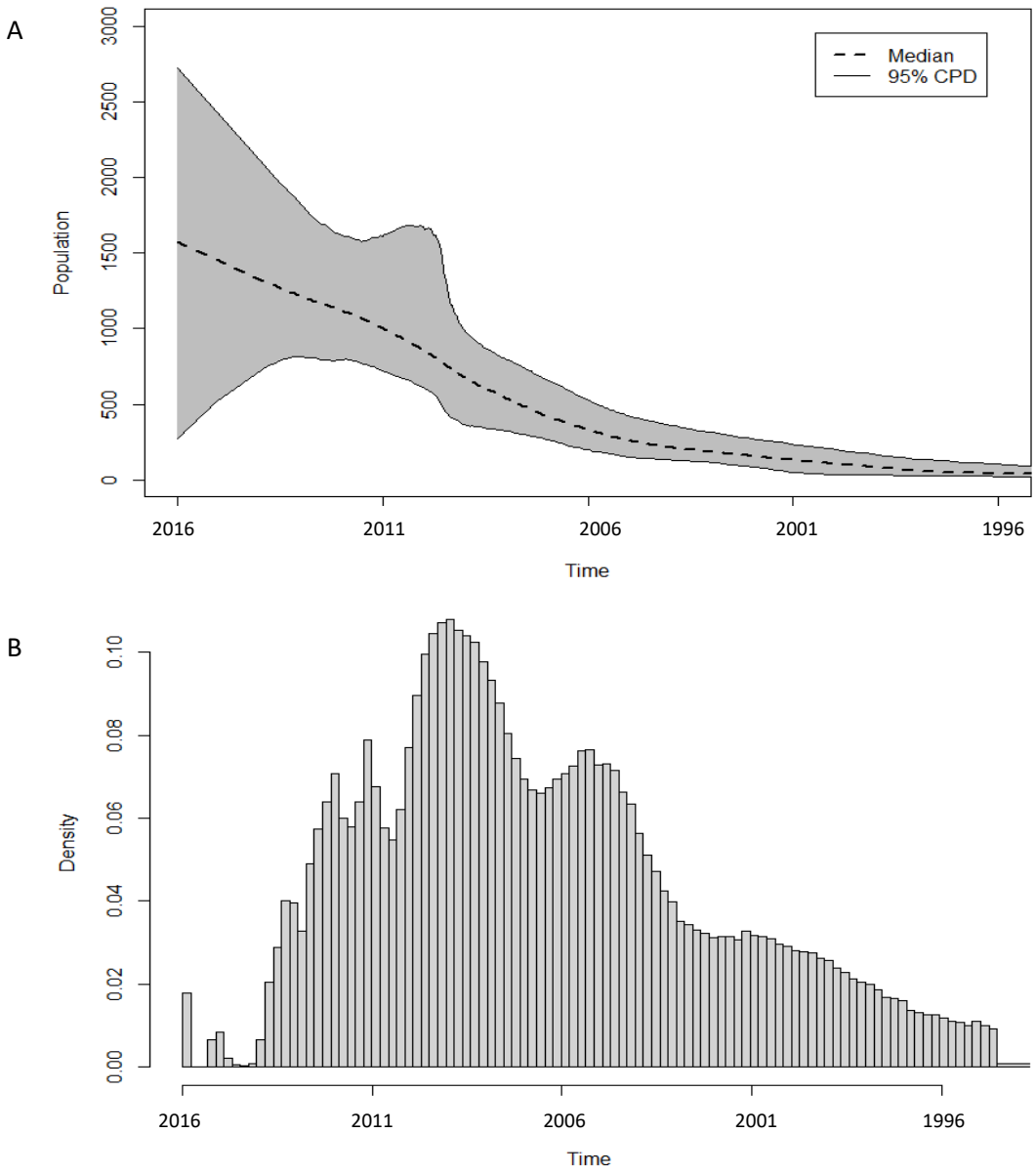

Figure 4.9. Population dynamics of S. flexneri 2a (A) and a histogram of tree events (B) through time.
Further into the past where there are fewer tree events, the prior information has a greater influence on the population size estimate. Time tree events correlate to the amount of influence the data has on the population size estimate.

4.4. Discussion

4.4.1. Genomic epidemiology of *S. flexneri* 2a cases shigellosis cases in South Africa

The generated global phylogeny confirmed the presence of endemic strains of *S. flexneri* 2a in South Africa. This study showed for the first time that multiple endemic strains of *S. flexneri* 2a are present across the whole country though distinct geographical distributions were observed for each of the identified BAPS clusters (Figure 4.8, Table 4.8). The interspersion of isolates from difference provinces within highly related clusters and across the tree suggests high levels of inter-provincial transmission (Figure 4.5). These findings agree with previous research into the *S. flexneri* which found strain persistence and co-existence, even following the introduction of new, successful strains into the region [33].

All *S. flexneri* 2a strains were identified as being part of Phylogroup 3 (Figure 4.5A), in agreement with the literature which has found serotype 2a predominantly in Phylogroup 3 [33]. Ancestral state dating estimated by my Bayesian tree model are extrapolated from a narrow time frame and may mean that our estimates further into the past are more inaccurate. A previous global *S. flexneri* phylogenetics study dated the MRCA of phylogroup 3 as much earlier, median 1848, than for my study isolates, the mid-1940s to mid-1960s (Figure 4.5C) [33]. However, as our study examines a national population it is expected that the MRCA would be more recent than a global Phylogroup estimate; geographical clustering of strains is common thus sampling only from South Africa would be unlikely to cover the full diversity of the known global *S. flexneri* 2a [33]. The identification of a more recent MRCA might also represent a national level introduction as for *Shigella sonnei* in Vietnam [73].

This study shows that the sub-cluster of Phylogroup 3 which contains all identified South African *S. flexneri* 2a is not exclusively found in South Africa (Figure 4.5A), though more research is needed to determine wider geographical distribution for this sub-cluster.

4.4.1.1. *Strains coexistence*

Most of the study isolates belonged to one of four endemic South African lineages (Figure 4.5A). However, some of the isolates which were grouped into BAPS3 are more closely related to reference isolates than the South African lineages and may represent imported strains (Figure 4.5A). The distribution of these likely-imported strains in the tree, and their grouping together with an endemic South African lineage within BAPS3, suggests high levels of transmission between the region of origin and South Africa. It seems likely that the region of origin for closely related strains would be a neighbouring African country, though little is known about the intracontinental transmission of *Shigella* strains within Africa. It is also possible that these strains have come from outside Africa as the international spread has been previously observed [78]. Overrepresentation of these probably imported isolates in Gauteng (58.8% or 10/17 isolates compared to 37.5 - 5.8% of isolates in endemic lineages) supports import from outside Africa as there is a large international airport province.

The four identified endemic South African *S. flexneri* 2a lineages showed variable AMR and virulence profiles (Table 4.5 and Table 4.7). One (BAPS2) was associated with MDR and systemic disease, another (BAPS1) with drug susceptibility and the presence of some virulence genes, while the remaining two (BAPS3 and BAPS4) were associated with MDR and a lack of systemic disease. The diversity of the identified endemic lineages confirms *S. flexneri* strain co-existence on a national level, something first suggested as part of a global study [33]. This work adds granularity to that picture showing that this occurs at the national, as well as regional, level.

This study found that the MDR lineages (BAPS2, 3 and 4) have a shared MRCA which was dated to the mid-1980s (Figure 4.5C). The MRCA of the drug susceptible population (BAPS1) was estimated to be younger than the MDR MRCA, likely emerging in the early 1990s (Figure 4.5C). This suggests MDR strains were already present in South Africa when the susceptible population emerged and found success.

A previous study of *S. flexneri* isolates from Somalia identified chromosome-based resistance to ampicillin, chloramphenicol, tetracycline and streptomycin, sampled in the mid-to-late 1980s, supports the widespread presence of MDR to these four antimicrobials across Africa at this time [185]. These isolates may share a common ancestor with our study isolates, or this may have arisen through convergent evolution driven by common antimicrobial used in these countries. Multiple, parallel acquisitions of the SRL, which encodes resistance determinants against these antimicrobials, have been previously observed [34].

The emergence-to-co-existence of a drug susceptible population alongside MDR strains suggests that these *S. flexneri* strains evolved and co-exist in different, but overlapping, ecological niches with distinct antimicrobial pressures. A large study into the evolutionary history of *S. flexneri* globally found drug susceptible clusters within Phylogroup 3, many isolates in these clusters were from Africa, but did not report the MRCA dating estimates for these clusters nor describe their probable evolutionary history, it is therefore not clear from the study if these clusters are co-existing or emergent [33]. While a study in Southeast Asia found no drug susceptible *S. flexneri* populations, the lack of susceptible isolates may reflect the *S. flexneri* population in Southeast Asia, rather than the global population, as there is strong evidence of high AMR selective pressures on *Shigella* in the region

The susceptible lineage (BAPS1) identified in this study may have increased virulence, based on the higher prevalence of virulence genes, which may have helped in its success (Table 4.7). The emergence of a pan-susceptible, highly virulent lineage has been observed in *Salmonella*, following a change in treatment regimen from chloramphenicol to ciprofloxacin in Africa [186]. This treatment change occurred between 2001 and 2003, after the emergence of this sub-population in South Africa. A similar dynamic, local to South Africa, may have happened here, however.

4.4.1.2. *Strain introduction*

The genetic distance between the susceptible lineage from the other endemic strains suggests this sub-population was recently introduced to the country (Figure 4.5C). Introduction and subsequent

successful colonisation of the country likely coincided with the estimated emergence date of this clade (the 1990s). The 1990's in South Africa saw the beginning of the HIV epidemic and the end of Apartheid.

Among the MDR cluster, one of the endemic lineages (BAPS4) is genetically distant from the other two endemic sub-populations (BAPS2 and 3), which are highly related (Figure 4.5C). This suggests the BAPS4 MDR sub-population is also a recent introduction to South Africa, likely introduced between 1999 and 2003. Likely aided by the increased cross-border travel post-Apartheid (Keddy, personal communication, see collaborator acknowledgements above). No obvious reason for the success of this recently introduced MDR population was identified in this study. However, the distinct geographical distributions of the different sub-populations may partially explain the successful introduction to co-existence of this lineage to the country; each sub-population dominating a different geographical niche would likely minimise direct competition.

Differing geographical niches are supported by minimal overlapping of high prevalence provinces between BAPS clusters. The only provinces strongly associated with more than one BAPS cluster were Gauteng and Northern Cape, provinces with the two largest cities in South Africa. These cities both have a high diversity of human sub-populations, each affected differently by socio-economic factors but living in proximity. Increased diversity of human sub-populations likely means increased diversity of pathogen ecological niches and may allow for multiple strain dominance. No province had isolates from only a single BAPS cluster, however, which is suggestive of coexistence. Little is known regarding local *S. flexneri* strain dynamics, more research is needed to understand the influence of inter-strain competition and the potential effects on strain epidemiology.

4.4.1.3. *Strain diversification*

Among the MDR sub-population, the closely related endemic BAPS2 and 3 lineages likely evolved from a single South African strain, diversifying into two distinct populations around 1991 (Figure 4.5C). While no factor influencing the success of the endemic BAPS3 lineage could be identified, the

emergence of BAPS2 may have been driven by an increase in virulence as BAPS2 is associated with systemic disease (though this association with systemic disease may be due to other unaccounted for e.g. socio-economic factors affecting patients). The role of non-genetic factors in the association between BAPS2 and systemic infection is supported by the lack of a virulence gene associated with the cluster; the different geographic distributions of the identified sub-populations, geography and socio-economic factors being highly correlated; and the lack of literature to support a link between systemic disease and pathogen factors.

The diversification of the South African MDR cluster and the likely introduction of two new successful endemic lineages confirm that multiple strains of *S. flexneri* can co-exist together. That the diversification and introduction events all likely occurred between the late 1980s and the early 2000s, coinciding HIV epidemic in South Africa (infection rate peaked 2000-2001) suggests that HIV has likely promoted diversifying evolution in the country (Thomas, personal communication, see collaborator acknowledgements above) [187].

The exact mechanism of HIV's influence on evolution is unclear. It may be that specific HIV associated lineages exist in the country, acting as a reservoir, or it may simply be that increased HIV prevalence in the country increased *Shigella* transmission and population size, leading to diversification. Without patient HIV status is it not possible to know if HIV-associated lineages exist. MSM-associated lineages have been previously identified however and some studies have found an association between shigellosis and HIV in the MSM community [188, 189]. The highest HIV prevalence, both during the study period and during the HIV epidemic peak of the 1990s, was in KwaZulu-Natal, however [150-154, 190, 191]. The lack of lineages associated with this high HIV prevalence province suggests that HIV-associated lineages may not exist or make up a small proportion of the population if they do exist.

The population dynamics modelling in this study suggest that the population size has been increasing since at least the early 2000s (Figure 4.9). This supports the role of population expansion in strain

diversification, likely linked to the HIV epidemic which has also seen an increase in prevalence since the early 1990s till the current day [187].

The association of the antimicrobial susceptible lineage (BAPS1) with Free State province provides support against the direct influence of HIV infection on the increased susceptibility in this lineage. However, HIV levels in Free State province may be under-reported due to the high number of mining communities which, in South Africa, often have private healthcare provided; patient data, such as HIV prevalence, is not always made publicly available (Thomas, personal communication, see collaborator acknowledgements above). Treatment regimens in these communities may be different to other regions which may have promoted the emergence of a susceptible lineage in the province. Alternatively, the emergence of the susceptible lineage could be due to socioeconomic factors specific to mining communities or to Free State province. Without further research, it is not possible to know what factors influence the success of this susceptible lineage.

4.4.2. Antimicrobial resistance

My results confirm widespread MDR in *S. flexneri* 2a across South Africa. Though emergence of a more drug susceptible lineage associated with Free State province may mean that treatment with antimicrobials (ampicillin, chloramphenicol, streptomycin and tetracycline) to which resistance is otherwise widespread would likely still be feasible, if only within Free State province.

Read mapping to the SRL shows that the difference in resistance between the susceptible lineage (BAPS1) and the other endemic lineages is likely due to the presence (MDR lineage)/absence (susceptible lineage) of the SRL, the breadth of mapping coverage for most MDR isolates being similar to that seen in the SRL positive control, around 90% (Figure 4.5C). Though other similar elements, such as the NR1 plasmid, have been previously identified and may be present rather than the SRL [82, 192]. The MDR sub-population isolates almost always have an SRL-like mobile element while the drug susceptible population isolates, generally, do not.

The phylogeny shows the loss and (re)acquisition of this element across all *S. flexneri* populations on a low level, though the prevalence in the respective sub-population appears stable (Figure 4.5C). Loss of the SRL, specifically or as part of the larger SRL-PAI, has been previously documented [193, 194]. Evidence of multiple acquisitions has also previously been found in *S. dysenteriae* [34]. It is therefore not unexpected for this locus to be inserted in isolation rather than as part of the SRL-PAI, as may have occurred in this study population.

The overall stability of the SRL-like element within each lineage, despite being generally unstable, points to the lineages existing in different ecological niches with distinct environmental conditions, one favouring presence of the SLR-like element and one favouring the absence. The environmental factors which may be influencing this are currently unknown. It is also possible they both currently exist in an ecological niche which has a completely neutral influence on the presence of the SRL.

4.4.3. Virulence

The presence of many known *Shigella* virulence genes, including complete virulence loci, were found to be almost ubiquitous in the *S. flexneri* population (Table 4.6). The presence of several complete virulence loci (SHI-1, SHI-2, *fim* and enterobactin) while lacking the *fec* locus in most isolates fits with previous findings on the virulence profiles of Phylogroup 3 isolates [33].

Minimal differences in virulence gene presence were observed between BAPS clusters, although, the drug susceptible lineage (BAPS1) was associated with the presence of two known likely-virulence genes, *capU* and *traT* (Figure 4.5C). Present across the population, *capU* was fixed in the susceptible lineage, while *traT*, generally rare, was common.

TraT, described as an outer membrane protein complement resistance gene in the virulencefinder database, may have a role in innate immunity evasion [195-197]. The function of *capU* has been less well studied; described as a hexosyltransferase homolog in the database based on protein sequence similarities, has been identified on a virulence plasmid (pAA2) associated with enteroaggregative *E. coli* [198, 199]. Very few studies have tried to determine the function or role of the capU protein,

though one showed that it was not likely to be involved in producing the biofilm phenotype associated with the pAA2 plasmid [198].

The presence of neither of these genes (*traT* nor *capU)* was associated with systemic infection and, thus, any increased virulence conferred by these genes is likely not related to bloodstream invasion. No direct link between a specific virulence gene and an invasive disease phenotype has yet been identified in *Shigella*. The factors influencing the development of systemic disease are not well understood and are likely predominately attributable to human host factors. The role for host factors in the development of systemic shigellosis is supported by the increased risk of systemic disease in HIV+ patients [200]. The population clustering of bloodstream invasion observed in this study could be suggestive of pathogen factor influence, though no specific factor was identified, nor can host or environmental factors be ruled out.

4.4.4. Conclusions

Multiple, distinct strains of *S. flexneri* 2a coexist in South Africa. Diversification since the late 1980s was likely promoted by the HIV epidemic in the country driving transmission and increasing population size. The HIV epidemic, alongside post-Apartheid cross-border migratory changes, may also have enabled the successful introduction of two new lineages into the country, one MDR and one drug susceptible, though there is no evidence for a direct influence from HIV on AMR acquisition or loss.

Three out of four endemic South African lineages are MDR, pointing to fitness benefits from maintain MDR. While the emergence of a drug susceptible lineage suggests that environmental pressures in the country are not exclusively MDR-promoting. Further research to understand these selection pressures may help reverse the increasing drug resistance seen in *Shigella* across the world.

Evidence of multiple conflicting environmental pressures is also evidence of *S. flexneri* occupying multiple, overlapping but distinct ecological niches. It is likely that other *S. flexneri* serotypes occupy the same niches as serotype 2a, given the similar epidemiology of *S. flexneri* serotypes [33]. Occupation of multiple niches would mean a multi-targeted approach is needed to reduce

transmission of *S. flexneri*. Further research to confirm the findings of this study and determine the forces promoting increased population size and diversification is needed.

Chapter 5

Genomic Epidemiology of Shigellosis in South Africa, part 2: *Shigella sonnei*

Contributions of collaborators

This chapter is the second half of the national study in South Africa, focusing on *S. sonnei*, the second most prevalent serotype across the study period. The contribution of my collaborators in this study are again outlined below.

Anthony Smith	Provided South African isolate metadata
Juno Thomas	Provided South African isolate metadata, official level of urbanisation classifications for South African provinces, and information on public healthcare policy and processes in South Africa
Karen Keddy	South African isolates sample selection and information on public healthcare policy and processes in South Africa
Neil Hall	Whole genome sequencing
Rebecca J. Bengtsson	Provided accession numbers for the *S. sonnei* phylogenetic reference isolates, having created known global population structure-representative, serogroup isolate lists, and smaller serogroup representative isolate lists. Also provided a working core-SNP alignment generation pipeline and bioinformatics support.

5.1. Introduction

Shigella sonnei is a highly prevalent global serogroup of *Shigella* that is the second most prevalent serotype in South Africa, accounting for around 30% of shigellosis cases [150-152]. Globally, S. *sonnei* has a distinct epidemiology from *S. flexneri* [33, 35]. The *S. sonnei* serogroup is younger and less genetically diverse than *S. flexneri*, having found global success through the international dissemination and clonal expansion of a single Clade [35]. The success of this Clade appears to be largely AMR driven [35, 39].

An association between level of urbanisation and the dominance of *S. sonnei* of a country has also been observed [35, 161]. It is possible that *S. sonnei* is more adapted to the environmental niches of industrialised areas compared to *S. flexneri,* having likely emerged in a highly industrialised region of the world (Europe during the industrial revolution), while the relatively older *S. flexneri* serogroup likely emerged in a pre-industrial world [33, 35]. One such adaptation could be cross-protection against *S. sonnei,* specifically, conferred by exposure to *Pleisiomonas shigelloides* due to similarity in the surface O-antigen of the two bacterial species [31, 75]. Other *Shigella* serotypes have variant surface O-antigens and likely lack cross-protection from *P. shigelloides* [71]. Exposure to *P. shigelloides* infections likely decreases with industrialisation due to improvements in access to clean water and sanitation and may account for the observed association between *S. sonnei* and industrialisation of a country.

It is also likely that *S. sonnei* is less adapted to pre-industrial environmental niches compared to *S. flexneri*. The deletion of toxin-antitoxin systems on the *S. sonnei* large virulence plasmid (pINV), required by all *Shigella* for successful infection, is known to have resulted in *S. sonnei* pINV instability, especially at environmental temperatures [24, 201]. Loss of the pINV outside of human hosts likely limits *S. sonnei* transmission via environmental passage, such as through contaminated water, while *S. flexneri*, having a stable pINV, would not be limited in this way and can likely take better advantage of the environmental transmission pathways more widely available in developing regions [24].

It is expected, due to the globally observed epidemiological differences, that the genomic epidemiology of *S. sonnei* on a national level will be distinct from that observed for *S. flexneri* in Chapter 4. In this study I used WGSA to examine the genomic epidemiology of *S. sonnei* in South Africa, describing the population according to the newly developed standardised nomenclature, modelling population and transmission dynamics, analysing the potential influences of AMR, and virulence characteristics of the identified strains on the observed epidemiology.

5.1.1. Aims

7. Describe *S. sonnei* epidemiology in South Africa
8. Identify endemic *S. sonnei* strain(s) in South Africa
9. Identify imported strains
10. Characterise the AMR profiles of all strains
11. Characterise the virulence gene profiles of all strains
12. Infer how strain characteristics influence the epidemiology

5.2. Methods

5.2.1. Selection and sequencing of South African *Shigella* study isolates in this study

The biochemically identified *S. sonnei* isolates (275 of 561 isolates in the sub-sample), collected as part of the public healthcare surveillance of *Shigella* in South Africa from 2011 to 2015, were introduced in the previous chapter (Chapter 4) and are included in this study. Collected as part of the same surveillance program as the *S. flexneri* isolates in the previous chapter, samples were representatively sub-sampled from the reported cases from across all nine provinces (Figure 5.2) across all five years. Samples were collected from patients, according to a standard diarrhoea definition, in a collection of public hospitals, from which they were sent to testing labs where they were biochemically characterised according to the surveillance laboratory Standard Operating Procedures (Supplementary) [150-154].

All *S. sonnei* isolates were subjected to the same whole genome sequencing and quality control as the South African *S. flexneri* isolates in the previous chapter using the methods laid out in sections 3.1 and 3.2 of the Methods chapter. The sequence quality of some of the original isolates was poor and so 65 of them were re-sequenced at the Centre for Genomic Research (CGR, University of Liverpool) using the Illumina NovaSeq 6000 platform; DNA library was constructed using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina [37]. The original and new sequence reads were combined for all re-sequenced isolates and then quality trimmed and assessed according to the standard protocol as laid out in sections 3.1 and 3.2 of the Methods chapter.

Of the selected *S. sonnei* isolates, no sequencing data was received for five, four were excluded due to uneven per base sequence content suggestive of a sequencing issue, and a further seven were excluded because the average genome mapping coverage depth was <20. Of the 259 which passed sequencing and mapping quality controls, six were excluded for not being *S. sonnei* (Supplementary), determined *in silico*, with shigaTyper (v1.0.6) and phylogenetic analysis (section 3.2 of the Methods chapter) [202]. The remaining 253 *S. sonnei* isolates were included in the study (57 were re-sequenced

isolates). All unpaired reads were also excluded from further analysis due to poor per base sequence content.

For some of the epidemiological analyses in this study were performed on the complete South African sample set of both *S. sonnei* and *S. flexneri* 2a. The same *S. flexneri* 2a sample set of isolates from the previous chapter was used for these comparisons, in this chapter. The complete sample set consisted of 513 isolates in total.

5.2.2. Sequencing of contextual study isolates in this study

A selection of previously sequenced reference isolates from across the known *S. sonnei* global phylogeny were also included for context (Section 2.1.4, Table 2.3) [4]. Phylogenetic reference isolate sequence reads were quality trimmed and assessed in the same way as the study isolates, see sections 3.1 and 3.2 of the Methods chapter.

5.2.3. Data collection

As for the South African *S. flexneri* 2a, recorded case numbers, extracted from the Group for Enteric, Respiratory and Meningeal Diseases Surveillance in South Africa (GERMS-SA) annual reports, were used to examine the background shigellosis epidemiology in the country, the larger sample set from which our study samples were selected, and the representation of that sample set by our study isolates [150-154]. Where possible, *S. sonnei* specific sample set comparisons were also carried out. Due to missing data in the 2014 and 2015 annual reports, comparisons were only able to be made using the 2011 to 2013 reported case numbers.

As in the previous chapter, the degree of urbanisation of district municipalities and metropolitan municipalities (both referred to as districts throughout) was defined according to the methods defined by the European Commission in March 2020 (https://ec.europa.eu/eurostat/cros/system/files/bg-item3j-recommendation-e.pdf). For this study, districts determined to be densely populated areas are

referred to as ‘urban’, intermediate density districts are referred to as ‘mixed’ and thinly populated districts as ‘rural’ (Thomas, personal communication, see acknowledgements above).

The association between patient age by gender was also examined. Patient date of birth and gender were recorded at the hospital at around the same time as sample collection, according to standard hospital procedures.

5.2.4. Population structure

The location of the *S. sonnei* study isolates within the known global *S. sonnei* population structure were determined with a *S. sonnei* maximum likelihood phylogeny as laid out in Section 3.2 of the Methods chapter, using a core-SNP alignment (6870 SNPS). The methods involved multiple software to map isolate quality trimmed sequence reads to, and then call variant sites against, a reference genome (*S. sonnei* strain 53G chromosome and plasmids: GenBank accessions HE616528, HE616529, HE616530, HE616531, HE616532), and define a consensus sequence. Gubbins (v2.3.4) was then used to define the core-SNPs, which were collected into a core-SNP alignment with SNP-sites (v2.4.1).

A selection of isolates from across the global serogroup phylogeny as previously determined were included in the phylogeny for global population structure reference [35]. The full list of included reference isolates and their accession numbers can be found in Section 3.2 of the Methods chapter.

Population structure modelling of just the South African *S. sonnei* isolates, in the context of time, was achieved with the second-generation Bayesian Evolutionary Analysis by Sampling Trees software (BEAST2) (v2.6.3) using the extended coalescent Bayesian skyline tree model [166]. Generated from a South African isolates only core-SNP alignment (4161 SNPs), re-created in the same way as for the maximum likelihood phylogeny without including the population structure reference isolates. The fasta formatted sequence alignment was converted to nex format with seqmagick (v0.8.0) (https://github.com/fhcrc/seqmagick/).

Two isolates (FD01874731, FD01874687), visually identified as having outlying, residual distances from the correlation between root-to-tip divergence and sampling date, were excluded from the BEAST phylogeny (Figure 5.1). Outliers have large residual distance from the correlation line of best fit as well as having log ancestor traces suggestive of a divergence time widely different from the other isolates (Figure 5.1). The correlation between root-to-tip divergence and sampling date, and isolate residuals, were visually assessed in TempEST (v.1.5.3) using a RAxML-ng (100 bootstrap, GTR+G model) generated maximum likelihood phylogeny from a *S. sonnei* study isolates-only core-SNP alignment (according to: http://beast.community/tempest_tutorial) (Figure 5.1) [165].

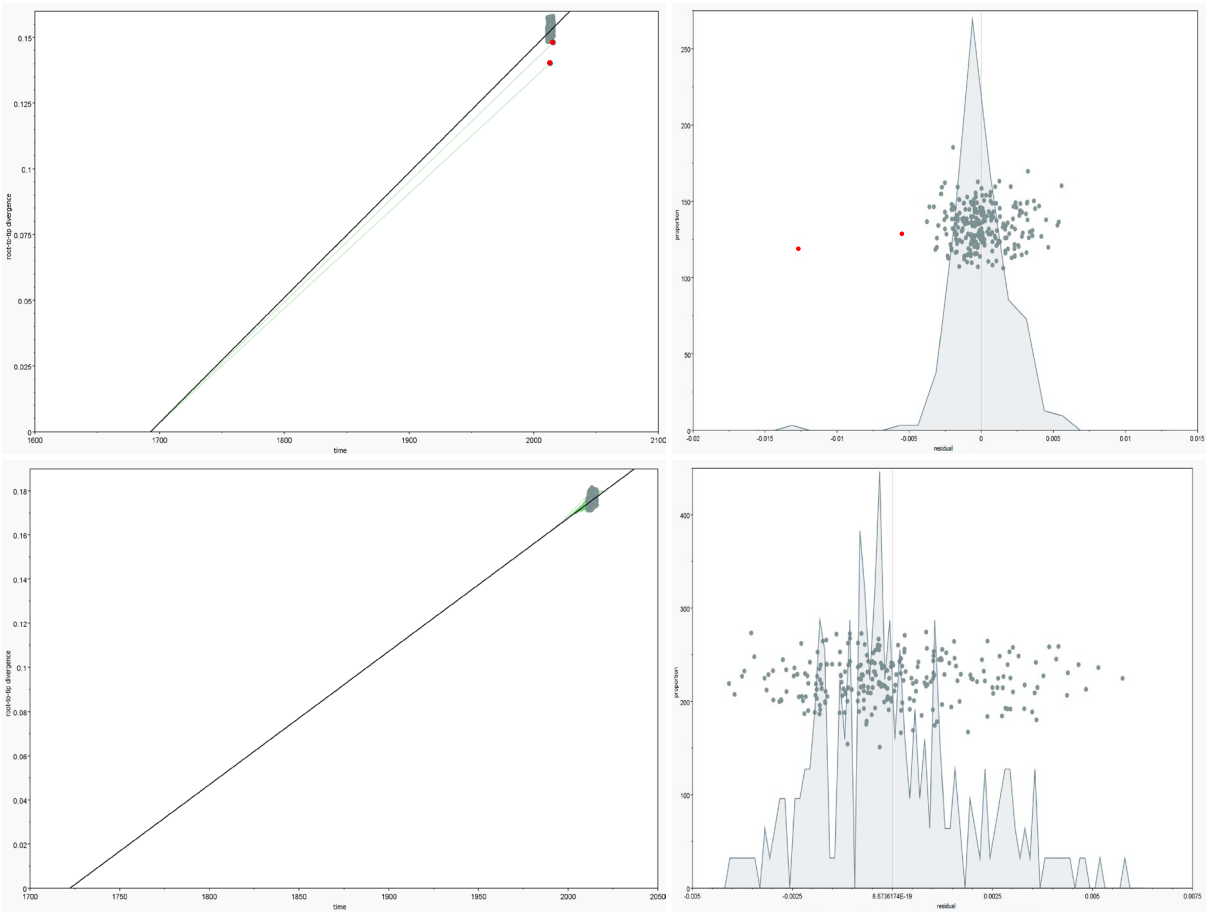

Figure 5.1. Correlation between root-to-tip divergence and of S. sonnei isolate sampling date (left) and the correlation residuals of these isolates (right) – before (top) and after (bottom) outlier isolates removed.
Excluded isolates marked in red, green lines show ancestor traces of excluded isolates (top left) and of all isolates (bottom left).

Inclusion of outlying isolates would obscure the molecular clock signal of the population, as it is an indication of sample date mislabling, contamination, sequence degradation or errors, or recombination. Once isolates excluded, the *S. sonnei* molecular clock rate was estimated to be $6.0361e^{-4}$, correlation co-efficient = 0.39, assessed in TempEST.

For the BEAST2 modelling, I used the same model parameters and priors as for the *S. flexneri* analyses in the previous chapter; a relaxed log normal clock model with prior rate = $1e^{-6}$ (single partition, based on https://beast2.blogs.auckland.ac.nz/tag/clock-rates), site model averaging with BmodelTest and transition-transversion split (prior mutation rate = 1.0, based on: https://beast2.blogs.auckland.ac.nz/tag/clock-rates), and all other priors left as default [167].

I ran three MCMC chains with a length of 3,300,000,000 (sampling every 100000). Around 67% burn-in was removed from all three runs while combining the output files with logcombiner. One of the three runs had a spike in the trace plot just before 67% into the run, so 67% burn-in was removed from the start of all three runs to ensure convergence. The tree topology was generated from the combined tree output files with treeannotator and then visualised in the interactive tree of life (ITOL) online platform and FigTree (v1.4.4) (https://github.com/rambaut/figtree), the latter was used to obtain node age estimates and 95% HPDs [158].

A minimum effective sample size (ESS) of 200 was achieved for all parameters once all runs completed, combined and burn-in removed, checked using Tracer (v2.6.3) [168]. The only exception to a minimum ESS of 200 was the bModelTest gamma shape parameter which had a very low ESS because the little support was found for site models with a gamma rate (has Gamma rates parameter median = 0). Sampling of gamma shape parameter only occurs when the model includes a site model with a gamma rate (has Gamma rates parameter = 1) and so sampling of was very low.

Population clusters of the South African isolates were defined using RhierBAPS (v1.1.3) a South African *S. sonnei*-only core-SNP alignment in Rstudio (v1.4.1717; R v4.1.0) [169, 170]. The output clustering from this is later referred to as BAPS clusters.

5.2.5. Strain typing

Alongside phylogenetic analysis, strains were typed according to a standard population structure-based typing scheme using the sonneiTyping sonnei_genotype.py and quality trimmed isolate reads (https://github.com/katholt/sonneityping).

5.2.6. Genome assembly

Draft genomes were assembled using Unicycler (v0.4.7) and quality assessed with Quast (v5.0.2), as laid out in section 3.4 of the Methods chapter [140, 149]. No isolates were excluded due to poor genome assembly.

5.2.7. Antimicrobial resistance profiling

Most isolates were phenotype tested against ampicillin (85% of isolates), chloramphenicol (64%), streptomycin (64%), tetracycline (64%), nalidixic acid (64%), ciprofloxacin (93%) and co-trimoxazole (64%) according to laboratory standard operating procedures (Supplementary). For isolates not phenotype tested, I used the same methods for antimicrobial resistance gene profiling and phenotype prediction for *S. sonnei* as used in the previous chapter for *S. flexneri*, as described in Section 3.4 of the Methods chapter. AMR genotyping was performed using the draft genome assemblies.

5.2.8. Virulence profiling

A virulence genotype profile was generated for all study isolates using the same methods as in the previous chapter; through comparison of isolate draft genome assemblies against the VirulenceFinder database, with VirulenceFinder (v2.0.4-1), and a local, curated virulence gene database (Table 5.1), with BLASTn (v2.10.0+) [155]. Only genes with ≥99% sequence identity and coverage were accepted. Virulence results from the previous chapter have also been included for comparison.

Table 5.1. Virulence genes included in curated database.

Gene	Accession number(s)
entA	NC_004337.2 (531914-532660), NZ_CP055292.1 (c618316-617570), NZ_MSJW02000146.1 (c17068-16322), NZ_LPTR01000087.1 (c7619-6873)
entB	NC_004337.2 (531057-531914), NZ_CP055292.1 (c619173-618316), NZ_LPTR01000087.1 (c8476-7619)
entC	NC_004337.2 (528248-529423), NZ_CP055292.1 (c621982-620807), NZ_MSJW02000146.1 (c20554-19379), NZ_LPTR01000087.1 (c11285-10110)
entD	NC_004337.2 (c512813-512184), NZ_CP055292.1 (636788-637408), NZ_LPTR01000126.1 (c1598-978)
entE	NC_004337.2 (529433-531043), NZ_CP055292.1 (c620797-619187), NZ_MSJW02000146.1 (c19369-17750), NZ_LPTR01000087.1 (c10100-8490)
entF	NC_004337.2 (516881-520726), NZ_CP055292.1 (c632710-628829), NZ_LPTR01000126.1 (5676-9557)
fecA	NZ_MSJW02000150.1 (2888-5212)
fecB	NZ_MSJW02000150.1 (5257-6159)
fecC	NZ_MSJW02000150.1 (6156-7154)
fecD	NZ_MSJW02000150.1 (7151-8107)
fecE	NZ_MSJW02000150.1 (8108-8875)
fecR	NZ_MSJW02000150.1 (1848-2801)
fepA	NC_004337.2 (c515219-512979), NZ_CP055292.1 (634373-636613), NZ_LPTR01000126.1 (c4013-1773)
fepB	NC_004337.2 (c527956-527000), NZ_CP055292.1 (622357-623313), NZ_MSJW02000146.1 (20928-21884), NZ_LPTR01000087.1 (11577-12533)
fepC	NZ_CP055292.1 (626668-627483), NZ_LPTR01000087.1 (15870-16685)
fepD	NC_004337.2 (c525665-524649), NZ_CP055292.1 (624678-625682), NZ_LPTR01000087.1 (13880-14884)
fepG	NC_004337.2 (c524652-523660), NZ_CP055292.1 (625679-626671), NZ_LPTR01000087.1 (14881-15873)
fimA	NC_004337.2 (c4382103-4381555), NZ_CP055292.1 (c1363528-1362980), NZ_CP055292.1 (4333156-4333719), NZ_CP055292.1 (c679609-679067), NZ_MSJW02000103.1 (c14019-13477), NZ_LPTR01000053.1 (7046-7594)
fimB	NC_004337.2 (c4384260-4383775), NZ_CP055292.1 (c1365685-1365083)
fimC	NC_004337.2 (c4380914-4380189), NZ_CP055292.1 (c1362339-1361614), NZ_CP055292.1 (c678847-678155), NZ_MSJW02000103.1 (c13257-12565), NZ_LPTR01000053.1 (8235-8960)
fimD	NC_008258.1 (c4350014-4347378), NZ_CP055292.1 (c1361547-1358911), NZ_LPTR01000053.1 (9027-11576)
fimE	NC_004337.2 (c4383180-4382584), NZ_CP055292.1 (c1364605-1364009), NZ_LPTR01000053.1 (5969-6565)
fimF	NZ_CP055292.1 (c1358901-1358371), NZ_LPTR01000053.1 (11586-12116)
fimG	NC_004337.2 (c4376157-4375654), NZ_CP055292.1 (c1358358-1357855), NZ_LPTR01000053.1 (12129-12632)
fimH	NC_004337.2 (c4375634-4374732), NZ_CP055292.1 (c675502-674495), NZ_CP055292.1 (c1357835-1356933), NZ_MSJW02000103.1 (c9912-8905), NZ_LPTR01000053.1 (12652-13554), NZ_LPTR01000196.1 (c1671-664)
iucA	NC_004337.2 (3820792-3822573)
iucB	NC_004337.2 (3822574-3823521)
iucC	NC_004337.2 (3823521-3825263)
iucD	NC_004337.2 (3825260-3826597)
iutA	NC_004337.2 (3826603-3828798)
pic	NC_004337.2 (c3071855-3067737)
sat	NZ_MSJW02000174.1 (772-4659), NZ_LPTR01000224.1 (c1528-95)
set1A	NC_004337.2 (3069744-3070277)
set1B	NC_004337.2 (3069555-3069740)
shiA	NC_004337.2 (3808392-3809435)
shiB	NC_004337.2 (3809950-3810411)
shiC	NC_004337.2 (3811744-3812262)
shiD	NC_004337.2 (3817102-3817500)
shiE	NC_004337.2 (3818599-3819570)
sigA	NC_004337.2 (3060437-3064294)
sitA	NC_004337.2 (c1408785-1407895), NZ_MSJW02000113.1 (20693-21607), NZ_LPTR01000169.1 (c5068-4154)
sitB	NC_004337.2 (c1407895-1407068), NZ_MSJW02000113.1 (21607-22434), NZ_LPTR01000169.1 (c4154-3327)
sitC	NC_004337.2 (c1407071-1406214), NZ_MSJW02000113.1 (22431-23288), NZ_LPTR01000169.1 (c3330-2473)
sitD	NC_004337.2 (c1406217-1405360), NZ_MSJW02000113.1 (23285-24142), NZ_LPTR01000169.1 (c2476-1619)
stx1A	NC_028685.1 (25526-26473), NC_025434.1 (20901-21848)
stx1B	NC_028685.1 (26483-26752), NC_025434.1 (21858-22127), NC_029120.1 (21872-22141)

The presence of the large virulence plasmid (pINV) was also assessed in the same way as in the previous chapter, for *S. flexneri* study isolates, except quality-trimmed isolate sequence reads were mapped to the *S. sonnei* 53G strain complete reference genome (Table 5.2). The presence of the pINV in this reference genome was visually confirmed in ACT following a BLASTn (v2.10.0+) comparison against a reference pINV sequence (Table 5.2). Read mapping, using BWA mem (v0.7.17), was assessed across the entire HE616529.1 plasmid [203, 204]. Samtools (v1.9) was used to remove unmapped, not primarily aligned, QC failed, duplicate, and supplementary reads, and sort and index the mapped reads [142, 144].

The read mapping of a positive and a negative control were also assessed (Table 5.2) to ensure that any observed read mapping was due to presence of the pINV and not an *in silico* read mapping artifact. Random reads (paired, 100-70 bases, insert size 150-300) were generated from the positive and negative controls (Table 5.2) with the bbmap randomreads.sh script (v38.00) to 60x coverage (www.sourceforge.net/projects/bbmap/). Converted into a set of forward and a set of reverse reads with seqkit (v0.10.1) [171].

Table 5.2. Reference sequences and positive and negative controls for assessing the presence or absence of virulence loci.

Virulence loci name	Reference loci sequence	Mapping reference genome	Positive control	Negative control
pINV	*S. flexneri* 5a M90T strain, pWR501 plasmid (NC_002698.1)	*S. sonnei* 53G strain, chromosome and plasmids (HE616528, HE616529, HE616530, HE616531, HE616532)	*S. sonnei* 53G strain, chromosome and plasmids (HE616528, HE616529, HE616530, HE616531, HE616532)	*S. sonnei* 53G strain, chromosome only (HE616528)

Presence or absence of the pINV was assessed using: 1) the mean read depth across the pINV, and 2) the breadth of coverage across the pINV, defined as the number of bases with a read depth >1 divided

by the total number of bases. Zero read mapping to the pINV was observed with the negative control, so absence was defined as mean read depth = 0 or the breadth of mapping coverage ≤5%.

Read depth at each site of the pINV was determined using bedtools (v2.29.2). The mean read depth, breadth of coverage and a graph of the read depth across the pINV for each isolate were defined using an in-house python script (Supplementary). Read mapping to the entire reference genome was also assessed, as laid out above. A minimum mean read depth ≥10 was required for an isolate to be included in the pINV mapping analysis, though no isolate failed to meet this requirement.

From the read depth graphs generated by the in-house read mapping summary stats python script (Supplementary), I defined a 'region of interest' on the pINV which had a much greater read depth than the rest of the plasmid in several isolates (Figure 5.2). To examine this further I assessed the mean read depth and breadth across this region. Having observed, in the Prokka (v1.14.6) generated annotation for the reference pINV (GenBank accession: NC_002698.1), that the region of interest was surrounded by insertion sequences, I defined the region as including the insertion sequences at both ends. This region of interest was defined as GenBank accession number: HE616529.1, Positions:76975-126397 (Figure 5.2). Read mapping across this region was assessed in the same way as laid out above for the entire pINV.

The virulence gene *capU*, identified in this population, has been previously found on the *E. coli* pAA2 plasmid, GenBank accession number: AF134403.1 [198]. To determine if *capU* was carried on a plasmid, all *capU* carrying contiguous sequences (contigs) were compared against the NCBI nt database with BLASTn (v2.10.0+). As the top hits from these comparisons were against *Shigella* and *E. coli* pINV sequences, all isolate contigs were compared against a reference pINV sequence (CP053752.1) with BLASTn.

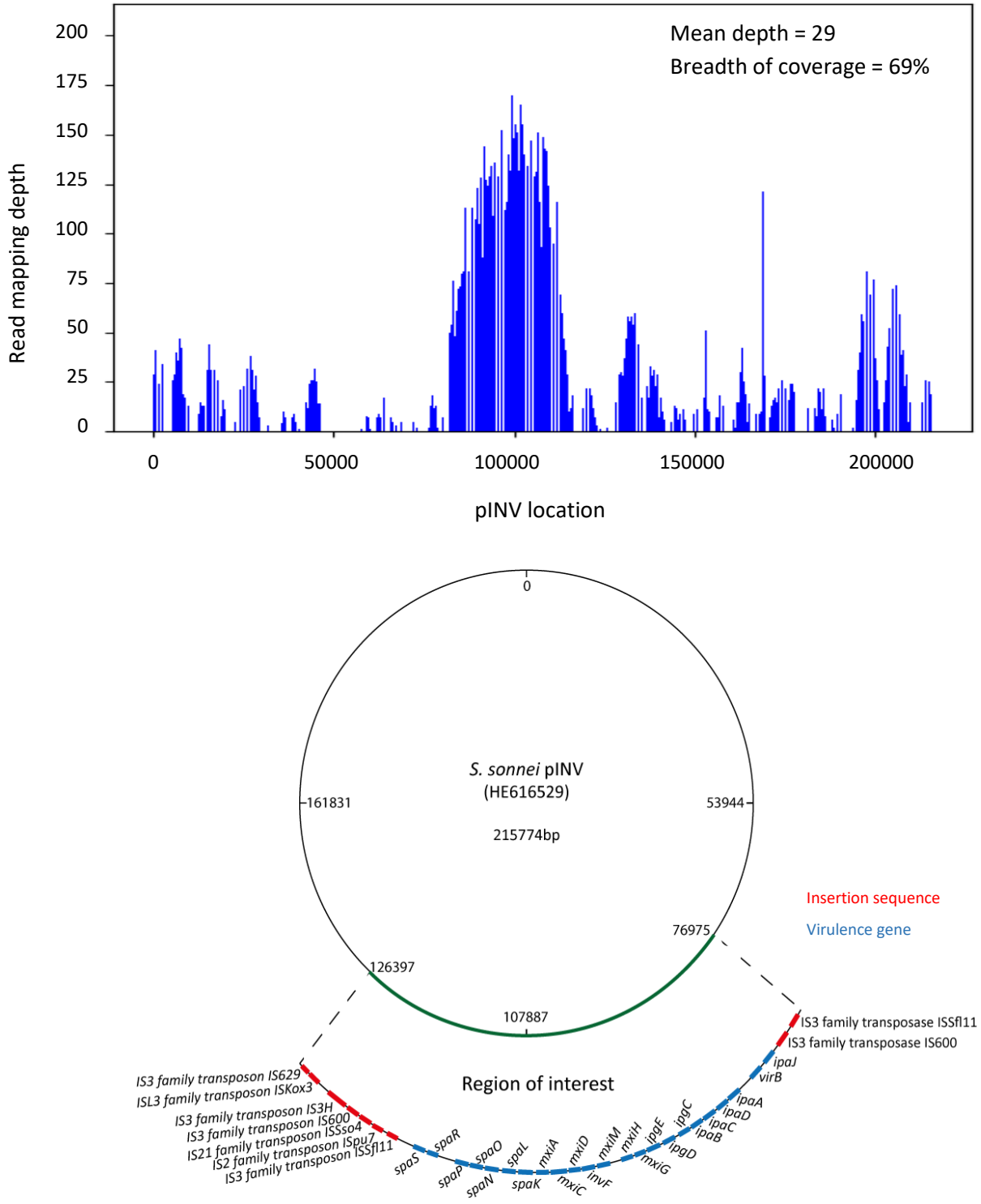

Figure 5.2. Example large virulence plasmid (pINV) read mapping graph (for study isolate FD01874140) (top) and a representation of the pINV 'region of interest' location and encoded genes (bottom).

5.2.9. Population dynamics

Dynamics of the South African *S. sonnei* population were also modelled during the BEAST2 population structure modelling (Section 5.2.3), with the extended coalescent Bayesian skyline model, in the same way as was done in the previous chapter for the *S. flexneri* 2a population (Chapter 4).

The Bayesian skyline model estimates the population size between multiple timepoint pairs across the modelled phylogeny. The plot of these estimated population sizes shows the changing population size through time. In the Extended version of this model, the optimum number of time points and the amount of time between them is estimated during the modelling process.

Two models of structured population dynamics were created, using marginal approximation of the structured coalescent (MASCOT). The populations were structured in two ways, one for each model: 1) patient sub-populations (based on age and gender) and 2) the level of urbanisation of the district of origin (rural, mixed and urban districts) (Table 5.3). MASCOT is a structured coalescent model which approximates rather than infers ancestral migration histories and thus requires less computing time, ideal for my large dataset.

Table 5.3. Table of structured coalescent model grouping and number of isolates per group.

Age and gender-based population stratification		**Level of urbanisation-based population stratification**	
Children <5 years old	n = 92	Rural	n = 24
Children ≥5 and <18 years old	n = 42	Mixed	n = 30
Adult women (≥18 years old)	n = 66	Urban	n = 182
Adult men (≥18 years old)	n = 51		

Where the same prior information was required in the structured coalescent models, I set the priors to have the same values as I used for the population structure and population dynamics model (Section 5.2.3). These priors were a relaxed log normal clock model, with a prior rate of 1e-6, and site

model averaging with bModelTest, prior rate 1.0, was used again for both structured coalescent models, and all other priors were set as default.

Patient age and gender metadata were not available for all isolates, nor was the specific sampling district. All isolates without the required metadata to assign a population structure category in the respective structured coalescent modelling were excluded (25% for age and gender analysis and 5% for level of urbanisation analysis).

For each model, three separate runs were used (sampling every 100000, for chain length 785,000,000), combined using logcombiner with 10% burn in removed from each while combining. Tracer was used to check model convergence, the ESS of the model parameters and determine the model estimates for parameters of interest (sub-population size estimates and between group migration rates). The only parameter which again did not have an ESS ≥200 was the bModelTest gamma shape parameter.

5.2.10. Statistics

The Chi-squared test of association was used to test most associations; carried out, in all cases, using the raw numbers but reported, in most cases, as a percentage difference. For the background epidemiology and sample set representativeness, these percentage differences are defined as the observed cases or samples (O) as a fraction of the national total (T) minus the expected (E) fraction of the national total ($\frac{O}{\mathrm{To}} - \frac{E}{\mathrm{Tc}}$), where T_o is the total observations (cases or samples) and T_c is the total reported cases. For associations between isolate or patient metadata associations, the percentage difference is defined as the difference between observed and expected number isolates as a fraction of the expected number of isolates ($\frac{O-E}{E}$).

Fisher's exact test was against used to test for association between AMR determinant or profile and geographic region. While odds ratios and two-tailed Z tests were used to assess the associations between AMR determinant or virulence gene and *Shigella* sub-population. The Bonferroni correction

was used to adjust the statistical significance threshold for multiple comparisons. The Pearson chi-squared test of independence was used to assess the association between pINV read mapping and presence of an AMR or virulence determinant.

5.3. Results

5.3.1. Epidemiology and evaluating representativeness of isolates

Assessing the levels of shigellosis across South Africa using the case reporting statistics from the Group for Enteric, Respiratory and Meningeal Diseases Surveillance in South Africa (GERMS-SA) annual reports revealed that some geographical areas had higher reported shigellosis cases relative to their population size [150-152]. Specifically, reported cases are higher than would be expected based on the proportion of the national population living in Gauteng (+14%) and Western Cape (+12%), while the converse is true of Limpopo (-10%), Mpumalanga (-5%), North West (-6%) and KwaZulu-Natal (-6%; $\chi^2(26, n=5078.01) = 2156, p=0.000$). This might represent an imperfect surveillance system or genuine differences in the incidence of the disease in different geographical areas.

To assess the representativeness of the sample set I compared the number of isolates to the expected number of isolates, based on the reported case numbers. The number of isolates in the complete sample set (both *S. sonnei* and *S. flexneri* samples) were representative of the reported case numbers (2011-2013) by both province ($\chi^2(26, n=353) = 19.02, p=0.890$) and year (2011-2013; $\chi^2(5, n=353) = 2.75, p=0.474$) (Figure 5.3A), indicating that the sample set was proportional to reported disease burden.

When stratifying this comparison by species and assessing the geographic associations of the *S. sonnei* isolates specifically, I found that the *S. sonnei* samples were associated with Gauteng (14% higher than expected, $\chi^2(8, n=253) = 35.69, p=0.0000$) (Figure 5.3B), reflecting the same association observed in the reported cases which show that 44-66% of *S. sonnei* cases are in Gauteng [150-152]. While no association between *S. sonnei* isolates numbers and year were observed. These results confirm the representativeness of our study isolates of the larger surveillance sample set, both as a complete, two serotype, sample set and as an *S. sonnei* specific sample set.

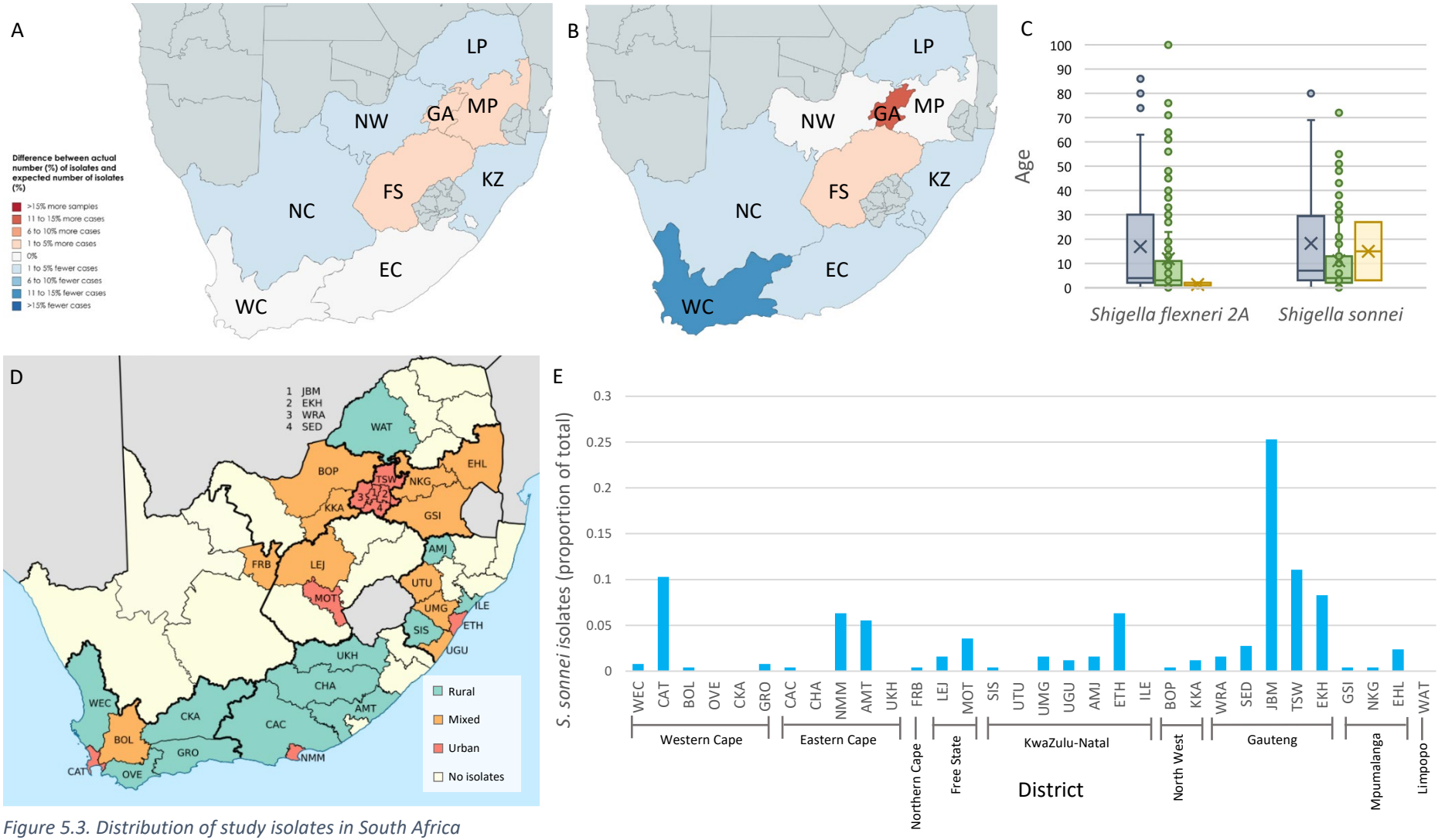

Figure 5.3. Distribution of study isolates in South Africa

A. Comparison of the number of isolates (*S. sonnei* and *S. flexneri*) against the expected number of isolates by province, based on the recorded number of cases 2011 to 2013. **B.** Difference between actual and expected number of *S. sonnei* isolates by province, based on recorded cases 2011 to 2013. **C.** Distribution of patient age by gender and serotype **D.** Urbanisation level of sampled district. E**.** Distribution of *S. sonnei* isolates across sampled districts, n = 253.

WEC = West Coast DM, CAT = City of Cape Town MM, BOL = Cape Winelands DM, OVE = Overberg DM, CKA = Central Karoo DM, GRO = Garden State DM, CAC = Sarah Baartman DM, CHA = Chris Hani DM, NMM = Nelson Mandela DM, AMT = Amathole DM, UKH = Joe Gqabi DM, FRB = Frances Baard DM, LEJ = Lejweleputswa DM, MOT = Manguang MM, SIS = Harry Gwala DM, UTU = uThukela DM, UMG = , UGU =Ugu DM , AMJ = Amajuba DM, ETH = City of eThekwini MM, ILE = iLembe DM, BOP = Bojanala Platinum DM, KKA = Dr Kenneth Kaunda MM, WRA = West Rand DM, SED =Sedibeng DM , JBM = City of Johannesburg MM, TSW = City of Tshwane MM, EKH = Ekurhuleni MM, GSI = Gert Sibande DM, NKG = Nkangala DM, EHL = Ehlanzeni DM, WAT = Waterberg DM.

The level of urbanisation of each district has been assessed according to the methods advised by the European Commission (https://ec.europa.eu/eurostat/cros/system/files/bg-item3j-recommendation-e.pdf). Most of the complete sample set (both serotypes) were isolates from urban districts (72%), while the fewest isolates came from mixed districts (9%) (Figure 5.3D and E). A likely consequence of the uneven surveillance across the country rather than shigellosis incidence.

Mixed and urban regions were both associated with *S. sonnei* (25 and 191/253 isolates, respectively), 10% and 4% higher, respectively, than would be expected if the number of isolates from each level of urbanisation were evenly distributed between the two serotypes ($\chi^2(2, n=513) = 7.489$, $p=0.0024$). This is likely reflective of a genuine stronger association between *S. sonnei* and increased urbanisation, when compared to *S. flexneri*. However, case numbers were not broken down by district in the surveillance reports, so I was unable to confirm a similar association in the reported case numbers.

When examining the distribution of patient age by gender, for the complete sample set (both serotypes) I observed no unevenness in the distribution of male and female patients between the two serotypes. I found that the average patient age was higher for females, median 5 years old (IQR = 2-30) than for males, median 3.5 years old (IQR = 1.25-11.75; Mann-Whitney U W= 18685, p=0.0027) (Figure 5.3C). This may be evidence of a genuine epidemiological phenomenon of older girls and adult women becoming infected at a higher rate than older boys and adult men.

The observed age difference in patients between men and women was more pronounced in *S. sonnei* relative to *S. flexneri 2a*, where the difference in median patient age was 3 years (median age female = 7 years old, median age male = 4 years old), than with *S. flexneri* 2a, where the difference in median patient age was 1 year (median age female = 4 years old, median age male = 3 years old) (Figure 5.3C). When stratifying patient age distribution by gender and serotype, the difference in age distribution between male and female patients was only found to be statistically significant in *S. sonnei* (Mann Whitney U test *S. sonnei:* W=5739.5, p=0.0072; *S. flexneri*: W= 7487, p= 0.0789). This could be reflective of a lack of an age of infection with *S. flexneri* infection or it could be because the effect is

smaller and the sample size of the study not being large enough to detect it. Either way this difference could be because of the increased stability of the *S. flexneri* pINV in the environment enabling environmental passage where *S. sonnei* might predominantly transmit through direct contact.

5.3.2. Population structure

To understand the evolutionary relationships of the study isolates within the known *S. sonnei* global population structure I generated a maximum likelihood phylogeny with reference isolates from across the known global phylogeny and the study isolates (Figure 5.4A). Most *S. sonnei* isolates (98.4%) were found to be part of the Global lineage III, forming a single main Clade (Global lineage 3.7) made of three sub-Clades (3.7.7, 3.7.9 and 3.7.11) and five more distantly related isolates (Figure 5.4A). The isolates not belonging to lineage 3.7 are diverse and include one lineage 5 isolate, a cluster of three lineage 2.8.2 isolates, two lineage 3.6 isolates and one lineage 3.4 isolate.

The sub-Clade nomenclature was defined according to the global *S. sonnei* standard with the *in silico* sonneiTyping software (Section 5.2.4). There was strong agreement between the sonneiTyping software and the maximum likelihood phylogeny, however, a small subcluster of the phylogenetically identified lineage 3.7 was predicted as Clade 3.7 but sub-Clade 2.8.2 by the sonneiTyping software, the single lineage 5 isolate was also misidentified as sub-Clade 2.8.2 (Figure 5.4A).

To examine associations between AMR, virulence factors, and geographic region with *S. sonnei* sub-populations the South African isolates were grouped into genetically related groups with BAPS. The BAPS population clustering grouped the Global lineage III isolates into three clusters (BAPS1-3), while all the lineage 2 isolates formed BAPS4 (Figure 5.4A). The three Global lineage III sub-clusters correlated strongly with the sub-clusters of 3.7 identified by the sonneiTyping software, BAPS1 = 3.7.11, BAPS2 = 3.7.7 and the 3.7 cluster misidentified as 2.8.2, and BAPS3 = 3.7.9. Though the BAPS2 sub-population also included the Clade 3.6 isolates, which are likely imported strains more closely related to strains from Central Asia [39]. The BAPS1 cluster was made up of all the correctly identified 2.8.2 isolates.

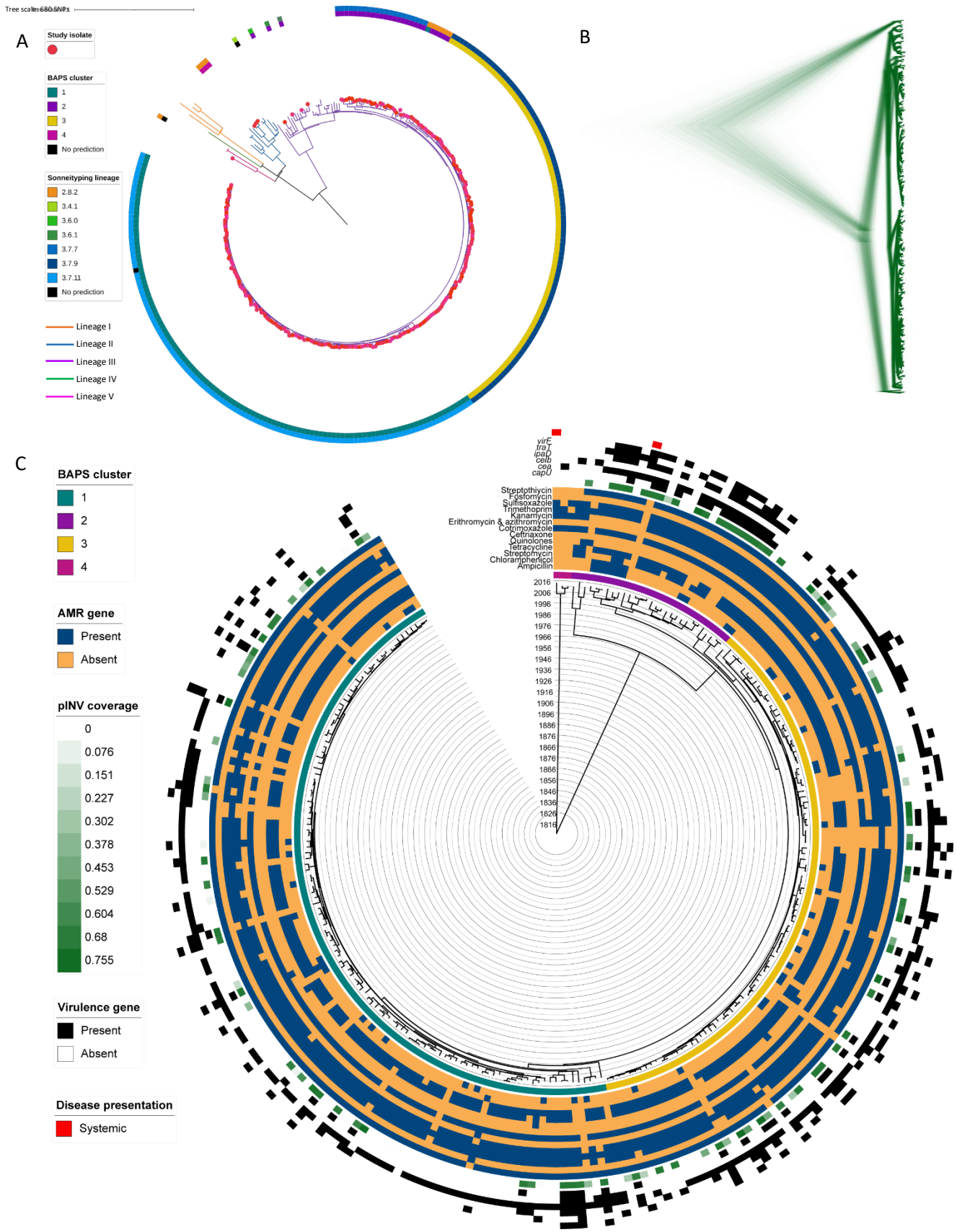

Figure 5.4. Population structure and clustering of predicted antimicrobial resistance phenotypes and virulence genotypes of S. sonnei.

A. Maximum likelihood phylogeny population structure of known global phylogroup 3, study isolates have pink terminal nodes. **B.** Statistical support for BEAST generated tree topology, visualised in densiTree. **C.** BEAST generated phylogeny, with branch length corresponding to time.

To define the evolution of *S. sonnei* in South Africa through time, a time-based phylogeny was created. This model found that the most recent common ancestor (MRCA) for the *S. sonnei* study isolates was dated to mid-1786 (95% HPD = mid-1714 to early 1849) (Figure 5.4C). The dating estimates from this phylogeny fit with previous estimates, providing support for the model [35]. The same model showed that the MRCA for all the endemic South African *S. sonnei*, coinciding with the MRCA for the BAPS2 cluster, was likely present in late 1967 (95% HPD = late 1957 to early 1978).

The BAPS1 and 3 sub-populations likely emerged more recently from the already present BAPS2 cluster in mid-1996 (95% HPD = mid-1993 to late 1999) and late 1997 (85% HPD = mid-1994 to early 2000), respectively (Figure 5.4C). However, when examining the likely emergence dates of the sub-populations defined according to the new nomenclature, all four sub-Clades likely emerged around the same time. The mis-identified 2.8.2 sub-Clade likely emerged in 1996 (95% HPD: 1993-1998), both sub-Clade 3.7.9 and 3.7.11 emergence estimates were in 1997 (95% HPD: 1993-1999 and 1994-2000, respectively), while sub-Clade 3.7.7 emergence was estimated in 1998 (95% HPD: 1996-2001). The emergence of these lineages appears to line up with the HIV epidemic peak in South Africa, anti-retroviral therapy for HIV was rolled out in 2004 in South Africa (Keddy, personal communications, see acknowledgements above).

5.3.3. Antimicrobial resistance

Antimicrobial resistance is an increasing hinderance to effective treatment of shigellosis. To characterise the type and level of drug resistance in *S. sonnei* in South Africa, I created genotypic AMR profiles and used these, in combination with partial phenotypic profiles, to predict AMR in the population. This showed that multidrug resistance (MDR), defined as resistance to three or more antimicrobial classes, was widespread (93.7% *S. sonnei* isolates) (Table 5.4 and Figure 5.5).

Little diversity in the identified resistance profiles was observed, 70% of isolates sharing the most common resistance profile: resistance to kanamycin, tetracycline, trimethoprim, streptomycin and sulfisoxazole (Figure 5.5B). However, twenty-eight isolates (11%) were predicted to be resistant to other antimicrobials, including ampicillin, chloramphenicol, azithromycin, cephalosporins, nalidixic acid and ciprofloxacin (Figure 5.5A). Showing that while most *S. sonnei* share the same AMR profile, or similar, resistance to a wide variety

of antimicrobials is present at a low level. This contrasts with those found for the *S. flexneri* 2a population in the previous chapter, only a minority of which shared the most common *S. flexneri* 2a AMR profile (38%) and only four isolates were predicted to be resistant to other antimicrobials outside those for which resistance to was highly prevalent (Table 5.4 and Figure 5.5).

Table 5.4. Antimicrobial resistance gene presence in the sample set, by serotype.
* = not associated with resistant phenotype, ¥ = point mutations, ^ only identified by AMRfinder

Genotype / phenotype	**Antimicrobial**	**Resistance determinant**	**Number of isolates (% of total isolates)**	
			S. flexneri 2a	*S. sonnei*
Phenotype and genotype	Ampicillin	*blaTEM-1A*	1 (0.4)	
		blaTEM-1B	32 (12.3)	17 (6.7)
		blaOXA-1	186 (71.5)	1 (0.4)
	Chloramphenicol	*catA1*	185 (71.2)	7 (2.8)
		cmlA1		1 (0.4)
	Tetracycline	*tet(A)*	11 (4.2)	198 (78.3)
		tet(B)	3 (1.2)	2 (0.8)
	Streptomycin	*aadA1*	195 (75)	239 (94.5)
		*aadA2**		
		aph(3'')-Ib	114 (43.8)	194 (76.7)
	Quinolones / fluoroquinolones¥	*gyrA* D87G		3 (0.12)
		gyrA D87Y		2 (0.8)
		gyrA S83L		3 (1.2)
		parC S80I		1 (0.4)
	Cephalosporins	*blaCMY-2*		1 (0.4)
		blaCMY-4		1 (0.4)
Genotype	Erythromycin and Azithromycin	*mph(A)*	3 (1.2)	2 (0.8)
		mef(B)		1 (0.4)
	Kanamycin	*aph(6)-Id*	159 (61.2)	221 (87.4)
	Gentamycin	*aac(6)-Ib-cr*	1 (0.4)	
	Trimethoprim	*dfrA1*	195 (75)	242 (95.7)
		dfrA5	2 (0.8)	
		dfrA7	1 (0.4)	1 (0.4)
		dfrA8	7 (2.7)	2 (0.8)
		dfrA12		1 (0.4)
		dfrA14	50 (19.2)	31 (12.3)
		dfrA15	2 (0.8)	1 (0.4)
	Sulfisoxazole	*sul1*	4 (1.5)	3 (1.2)
		sul2	159 (61.2)	1 (0.4)
		sul3		1 (0.4)
	Streptothricin^	*sat2*	195 (75.0)	241 (95.3)
	Fosfomycin	*fosA3*		1 (0.4)
	Macrolide-lincosamide-streptogramin B^	*ermD*	249 (96.8)	

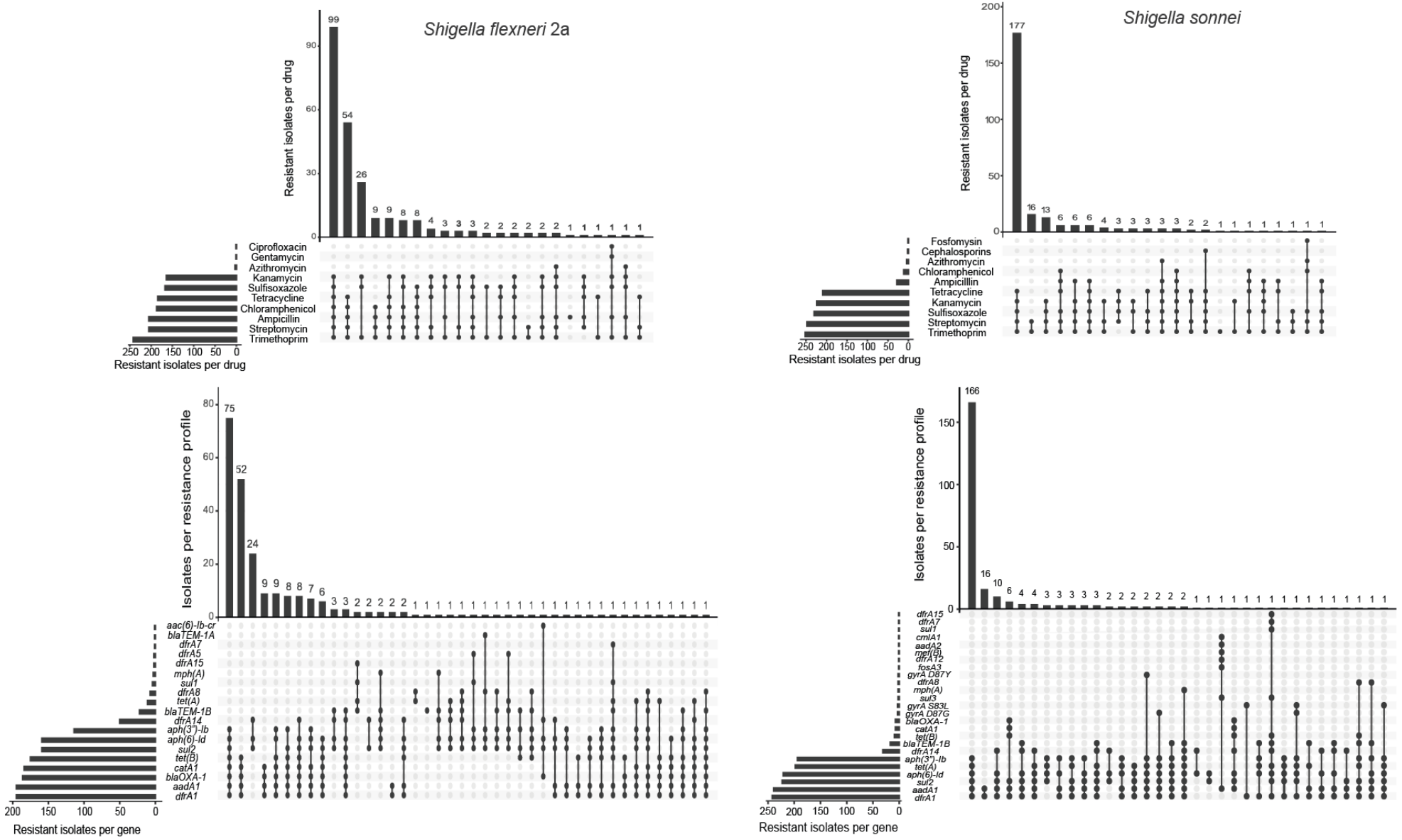

Figure 5.5. Antimicrobial resistance profiles in Shigella.

A. Predicted phenotypic profiles in *S. flexneri*. **B.** Predicted phenotypic profiles in *S. sonnei*. **C.** Genotypic profiles in *S. flexneri*. **D.** Genotypic profiles in *S. sonnei*.

Resistance to two remaining widely effective antimicrobials fluoroquinolones and third generation cephalosporins was identified at a low-level in *S. sonnei* (Table 5.4). Resistance to cephalosporins was phenotypically identified in two isolates, likely conferred by *blaCMY-2* and *blaCMY-4*. However, the only predicted FQR *S. sonnei* isolate, based on triple QRDR mutation, was reported as phenotypically susceptible to ciprofloxacin.

Point mutation based FQR is multi-step. A single point mutation in *gyrA* can confer intermediate resistance to FQR and full resistance to quinolones. In all four phenotypically identified quinolone resistant *S. sonnei* isolates, a QRDR point mutation was detected (Table 5.4). Two of the predicted quinolone resistant isolates were lineage 3.6.1 isolates, a subclade more often associated with Asia, and so were likely imported (Figure 5.4).

Some isolates were predicted to be pan-susceptible though this was rare (8/253 *S. sonnei* isolates). Pan-susceptibility was found to be associated with Mpumalanga province (1/8 isolates, 1481% greater than expected, fisher's exact two-tailed p=0.0005), an intermediate density province.

5.3.4. Virulence

To see if virulence was involved in the shaping the epidemiology of *Shigella*, genotypic virulence profiles were generated for the study isolates. A wide variety of virulence genes were identified in the *S. sonnei* population, however, many of the assessed virulence loci were only partially present (Table 5.5). This contrasts with the *S. flexneri* study population, in which the prevalence of these loci was high (Table 5.5).

Table 5.5. Virulence gene prevalence by serotype

Virulence Locus	Virulence gene	Number of isolates (% of total isolates)	
		S. flexneri 2a	*S. sonnei*
SHI-1	*pic*	260 (100.0)	155 (61.3)
	sigA	260 (100.0)	253 (100.0)
SHI-2	*shiA*	260 (100.0)	252 (99.6)
	shiB	248 (95.4)	
	shiC	256 (98.5)	
	shiD	213 (81.9)	
	shiE	260 (100.0)	253 (100.0)
	iucA	259 (99.6)	253 (100.0)
	iucB	259 (99.6)	253 (100.0)
	iucC	259 (99.6)	253 (100.0)
	iucD	259 (99.6)	253 (100.0)
	iutA	259 (99.6)	253 (100.0)
Enterobactin	*entA*	229 (88.1)	253 (100.0)
	entB	229 (88.1)	253 (100.0)
	entC	229 (88.1)	253 (100.0)
	entD	231 (88.8)	
	entE	229 (88.1)	
	entF	235 (90.4)	2 (0.8)
	fepA	235 (90.4)	249 (98.4)
	fepB	229 (88.1)	253 (100.0)
	fepC	235 (90.4)	253 (100.0)
	fepD	235 (90.4)	2 (0.8)
	fepG	235 (90.4)	5 (2.0)
fec	*fecA*		251 (99.2)
	fecB		250 (98.8)
	fecC		250 (98.8)
	fecD		250 (98.8)
	fecE		251 (99.2)
	fecR		251 (99.2)
Fimbriae	*fimA*	260 (100.0)	167 (66.0)
	fimB	260 (100.0)	
	fimC	260 (100.0)	
	fimD	260 (100.0)	4 (1.6)
	fimE	260 (100.0)	
	fimF	260 (100.0)	
	fimG	260 (100.0)	4 (1.6)
	fimH	260 (100.0)	
	capU	154 (59.2)	53 (20.9)
	cea		40 (15.8)
	celb	3 (1.2)	161 (63.6)
	cia		19 (7.5)
	cib		13 (5.1)
	gad	77 (29.6)	13 (5.1)
	ipaD	253 (97.3)	72 (28.5)
	lpfA	207 (79.6)	153 (60.5)
	mcbA	1 (0.4)	1 (0.4)
	ompT	1 (0.4)	
	astA		1 (0.4)
	sat	259 (99.6)	2 (0.8)
	senB		251 (99.2)
	virF	260 (100.0)	53 (20.9)

Both *Shigella* virulence loci (SHI-1 and SHI-2), present in nearly all *S. flexneri,* were partially present in *S. sonnei*; 61.3% isolates having both SHI-1 virulence genes and 6/10 SHI-2 genes identified (Table 5.5). While 100% of *S. flexneri* had all the fimbriae genes, only three were found in the *S. sonnei* population. Similarly, only half of genes in the enterobactin gene cluster were highly prevalent in *S. sonnei* while 88% of *S. flexneri* isolates had all of them. Conversely, though, none of the genes typically found in the *fec* locus were identified in the *S. flexneri* isolates but were all present in 99% of *S sonnei* isolates (Table 5.5).

The pINV is a large virulence plasmid unique to *Shigella* and EIEC which encodes the necessary genes for their enteroinvasive pathogenesis. Using read mapping methods to assess the presence or absence of the pINV, I found 59% of isolates lacked a pINV with only 42% of *S. sonnei* isolates having read mapping breadth greater than 20% (Figure 5.4C), mean read depth was <10 for many of these isolates.

Of those where mean read mapping was ≥10 (25% of isolates) the mean breadth of coverage was 69%. This suggests that, at most, only 25% of isolates had a complete pINV present, supported by the positive control which showed a similar level of mapping (mean depth = 41 reads, breadth = 73%) (Figure 5.4C). At least 16% of isolates had a partial/variant pINV or have no pINV but chromosomally located pINV sequences (Figure 5.4C). This is supported by the lack of mapping observed with the negative control, suggesting that observed mapping is not an artifact generated *in silico*.

Owing to the sub-population associations, discussed below (Section 5.3.5), that were found for *capU* and the previous identification of this gene on a plasmid (pAA2) in the literature [198], a BLASTn comparison of *capU* encoding isolate contiguous sequences (contigs.) with the NCBI nt database was undertaken. Common top hits of the *capU* containing isolate contigs were to pINV sequences from both *E. coli* and *Shigella* (data not shown), suggesting that *capU* is carried on the pINV in some of this *S. sonnei* population.

To see if other pINV contigs could be identified, a BLAST comparison of all isolate contigs against a reference pINV was performed. For those isolates with hits against the pINV, the mean number of hits against the pINV was 2.5 times greater for *capU*-carrying isolates than *capU*-lacking isolates (mean = 7.5 hits, vs mean = 3). Hits were defined as having an e-value of 0.0 and were between 486089 and 200 bases in length, suggesting that there are at least two different pINV versions present in the population, one with *capU* and one without. Though it may also be possible that the *capU*-lacking isolates have dropped their pINV plasmids and that their contig hits against the pINV plasmid are due to chromosomally located pINV sequences.

An association between the presence of *capU* and pINV mapping breadth was also observed (Pearson $\chi^2(1, n=253) = 120.36$, $p=0.000$), mean breadth of mapping across the pINV in *capU* carrying isolates was 69% compared to 14% in those without *capU*. Of the *capU* carrying isolates, 79% (42/53) had mean depth ≥10, though not every isolate with mean mapping depth ≥10 across the pINV carried *capU,* however, as 16 isolates (mean breadth of coverage = 66%) did not. These results provide further evidence of multiple pINV variants within the population.

I observed that read mapping to a 'region of interest' in the pINV had a greater average depth than the pINV, as a whole (see Section 5.2.6), suggestive of duplication of the region (Figure 5.6). This region encodes the mix-spa virulence locus as well as several other pINV associated virulence genes and starts and ends with multiple insertion sequences (Figure 5.2). When comparing the mean mapping depth across this region to the mean mapping depth across the pINV, many isolates had approximately double or triple (or more) coverage (Figure 5.5). No association between 'region of interest' duplications and system disease as observed.

For most of the isolates with 2 or more times the mapping depth, the breadth of mapping coverage across the pINV was approximately 70% (Figure 5.5). In these cases, it is likely that there has been a duplication event, one or more time, of the virulence associated region. For others, however, it was as low as 20% and in these cases, it may be that deletions of other regions of the pINV have brought

the total average mapping depth down relative to the mapping across the virulence-associated region, rather than a duplication.

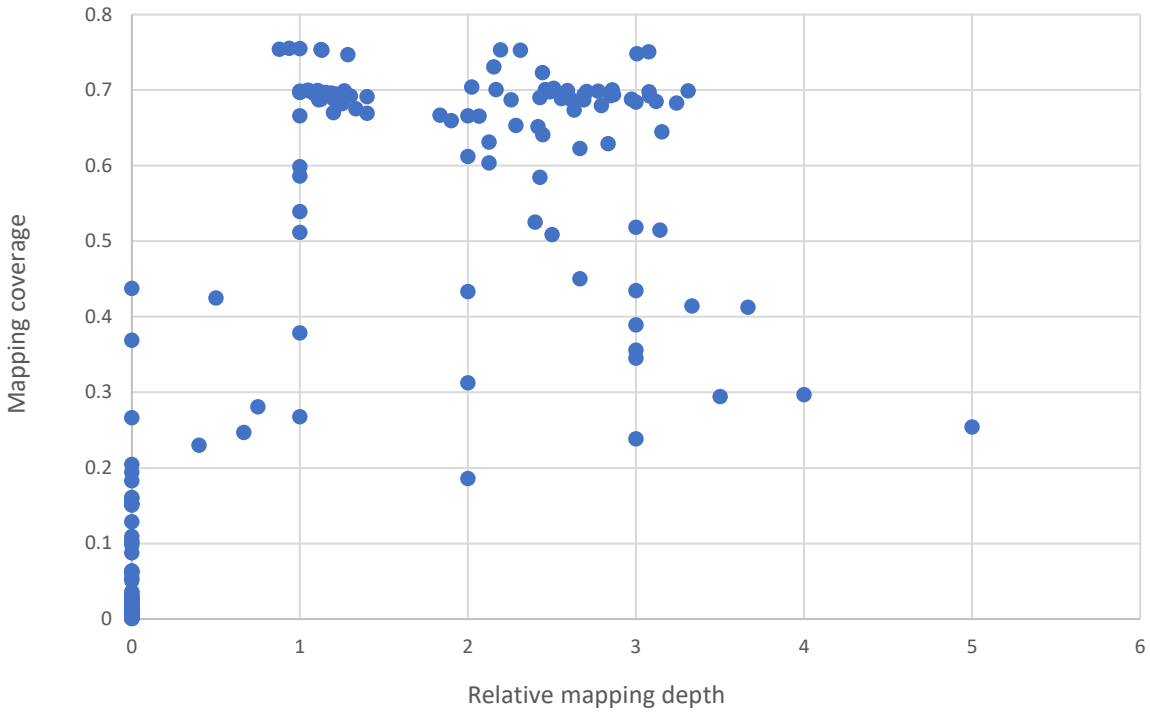

Figure 5.6. The mapping depth across the virulence-associated region of interest in the pINV relative to the mapping depth across the whole pINV, compared to the breadth of mapping coverage across the pINV.

5.3.5. Sub-population resistance, virulence and geographic associations

To characterise the identified sub-populations, a series of statistical tests were performed to see if statistically corrected associations with specific AMR or virulence determinants, region, or level of urbanisation and any sub-population existed. No associations with antimicrobial resistance, province, or level of urbanisation were observed. Both quinolone resistance conferring mutations and cephalosporin resistance genes were dispersed through the population.

Several virulence genes did, however, show cluster associations (Table 5.6). Most virulence genes identified as showing population clustering were positively associated with BAPS2 and negatively associated with BAPS1. An association was also observed between BAPS2 and having a mean pINV read mapping depth ≥10 (Pearson $\chi^2(1, n=253) = 6.919$, $p=0.009$) (Figure 5.3C).

Table 5.6. Associations between virulence gene and phylogenetic cluster in S. sonnei.
P-values calculated with two-tailed z test, significance threshold: p < 0.00143.

Gene	S. sonnei BAPS cluster	Proportion of cluster with gene	Odds ratio			P-value		
			Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
capU	1	0.13	1					
	2	0.48	**0.27 (0.67-0.11)**	1		0.0000		
	3	0.23	0.58 (1.17-0.28)	2.13 (5.08-0.89)	1	-	-	
	4	0	-	-	-	-	-	-
cea	1	0.14	1					
	2	0.69	**0.2 (0.52-0.08)**	1		0.0000		
	3	0.01	**13.52 (103.51-1.77)**	**66.9 (558.13-8.02)**	1	0.0006	0.0000	
	4	0.33	0.42 (3.20-0.05)	2.07 (25.87-0.17)	**0.03 (0.69-0.00)**	**-**	-	0.0001
celb	1	0.67	1					
	2	0.07	**9.75 (43.04-2.21)**	1		0.0000		
	3	0.78		**0.09 (0.40-0.02)**	1	-	0.0000	
	4	0	-	-	-	-	-	-
ipaD	1	0.19	1					
	2	0.59	**0.32 (0.77-0.14)**	1		0.0000		
	3	0.32	0.32 (0.60-0.17)	1.83 (4.31-0.78)	1	-	-	
	4	0	-	-	-	-	-	-
traT	1	0.11	1					
	2	0.38	**0.28 (0.72-0.11)**	1		0.0003		
	3	0.16	0.28 (0.62-0.13)	2.3 (5.78-0.91)	1	-	-	
	4	0	-	-	-	-	-	-
virF	1	0.1	1					
	2	0.52	**0.19 (0.49-0.07)**	1		0.0000		
	3	0.34	0.19 (0.44-0.08)	**3.86 (9.82-1.52)**	1	-	0.0000	
	4	0	-	-	-	-	-	-

While the number of isolates from patients who presented with invasive disease is too small for testing associations statistically, one of the two was a BAPS2 isolate, both BAPS1 and 3 lacked any invasive isolates despite having a greater number of isolates. The other invasive isolate was part of lineage 2.8.2 and was likely imported.

Genome recombination, detected with gubbins, was identified between bases 1310814-1316118 of *S. sonnei* 53G 301 strain chromosome was associated with lineage 3.7.11 (BAPS1). A region containing a sigma-dependant operon (*yciE*, *yciF*, *yciG*) and part of the tryptophan operon (*trpA*, *trpB*, and *trpC*). The tryptophan operon encodes enzymatic subunits involved in the synthesis of tryptophan. Little is known of the other genes in the recombinant region; however, the sigma regulon is stress induced [205].

5.3.6. Population dynamics

To examine how the population size changed over time, I used coalescent Bayesian skyline phylogenetic modelling to estimate the population dynamics. The output plot of estimated population size through time shows the population size likely rose sharply from around 2008-2011 till at least 2013 (Figure 5.6).

Post-2013 the level of uncertainty increases due to the reduced number of coalescence events and therefore reduced amount of input data to estimate from (Figure 5.6). The output dynamics graph shows a high level of certainty further into the past, however, the number of coalescence events pre-2006 are again very few, the level of certainty in this case comes from estimating from the prior information input into the model (Figure 5.6). The accuracy of the population dynamics post-2006, though, may not be very high.

These results suggest that the population size went through a large increase 2007 to 2013, though what happened before and after this is unknown. The estimated population increase could be a genuine increase in population size, or it could be the result of strain replacement, where the

previously dominant strain is completely undetected in this study and the population size erroneously low further into the past.

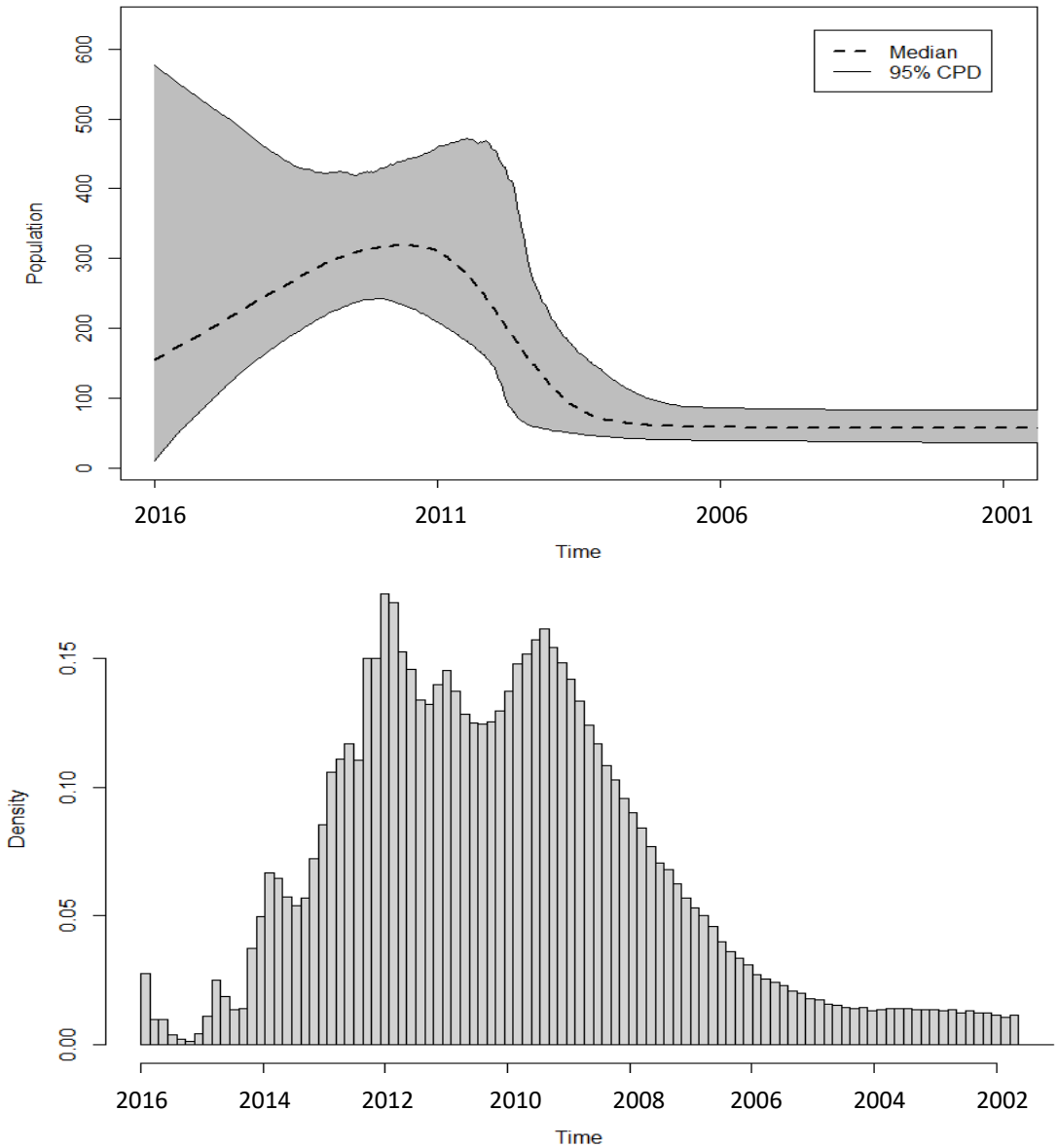

Figure 5.7. Population dynamics of S. sonnei (top) and a histogram of tree events through time (bottom). Time is time into the past and time point 0 = 2016. Further into the past where there are fewer tree events, the prior information has a greater influence on the population size estimate. Time tree events correlate to the amount of influence the data has on the population size estimate.

As I found a difference in the average age of infection between men and women (Section 5.3.1), I wanted to examine the differences in transmission rates between different age and gender groups.

To do so I used Bayesian structured coalescent phylogenetic modelling, structuring the population on the bases of age and gender.

This model estimates the population sizes for each group. While no statistically significant difference in population size was found, possibly due to the level of uncertainty leading to overlapping 95% HPDs, we know from the literature and surveillance that the age group with the greatest proportion of *Shigella* infections are those under the age of five years.

My model found that young children and adult women (young children: 124, 95% HPD: 29-289 and women: 143, 95% HPD: 12-351) had the largest median bacterial population sizes. The greatest uncertainty of the population size of any group is that of adult women, which may explain why it has a higher median estimate population size than under-fives. Adult men had the smallest median population size (15, 95% HPD: 4-34) and older children the second smallest (49, 95% HPD: 8-175).

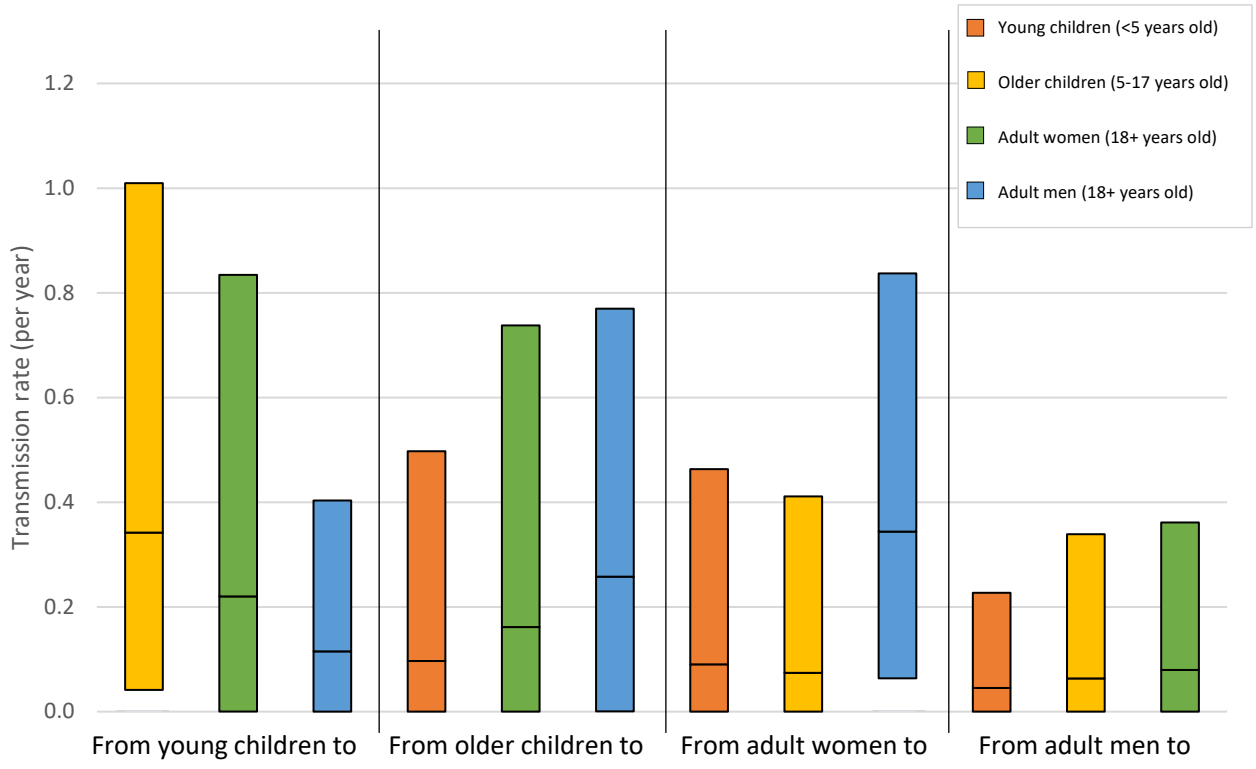

Figure 5.8. Estimated transmission rates (median ±95% HPD) from host sub-populations (x-axis) to other host sub populations (data, grouped by age and gender, coloured according to the inlaid key).

No statistically significant differences in the migration rates were detected by the model either. However, the estimated migration rates were consistently low from adult men to the other groups, suggesting that they may be the group contributing least to transmission (Figure 5.8). The greatest estimates of transmission rates to adult men are from adult women and older children.

Adult women may not contribute a large amount to transmission, except transmission to adult men. While children, younger and older, may contribute to a large proportion of overall transmission, particularly younger children which have high transmission estimates to older children and adult women (Figure 5.7).

While there is a lack of statistical significance, likely due to the level uncertainty in the model, these results suggest adult men are most likely to be infected by adult women, who are primarily infected by children under the age of five years old, though transmission likely occurs between all groups (Figure 5.8).

5.4. Discussion

5.4.1. Epidemiology

This study confirmed the presence of *S. sonnei* across South Africa, identifying multiple highly related, endemic sub-Clades of *S. sonnei* circulating in the country as well as evidence of strain importation from around the world.

An association between *S. sonnei* and Gauteng province, relative to *S. flexneri*, was observed in both the sample set and reported cases, indicating a genuine association and not selection bias. It is not known why this association exists; however, Gauteng is the most densely populated province in South Africa, every district in the province is defined as a densely populated region according to the level of urbanisation classification methods laid out by the European Commission. It is possible that the population density, and other socioeconomic factors, play a role in the dominance of *S. sonnei* in the region.

The role of urbanisation is supported by the association observed between *S. sonnei* and more urbanised districts compared to *S. flexneri* though it is not possible to rule out urbanisation as a confounding factor. Previous studies have observed an association between *S. sonnei* and industrialisation of a country, providing further support [35, 161].

The observed geographical distribution (Figure 5.3B) appears to be shared by all the identified endemic *S. sonnei* sub-populations. This suggests that all the identified sub-populations are likely occupying the same ecological niche, are likely subject to the same selection pressures, and are in direct competition with each other.

5.4.1.1. *Endemic sub-populations*

Four endemic sub-populations were identified, along with evidence of imported strains. All the endemic strains belong to Clade 3.7, previously known as the Global Lineage III clade as it has been highly successful at disseminating around the world [35, 39]. While the endemic South African *S.*

sonnei population is made up of multiple sub-Clades they are likely highly related, forming neighbouring clusters both in this study and in previous work [39].

All three of the endemic sub-populations from previously identified sub-Clades (3.7.7, 3.7.9 and 3.7.11) have previously been identified only in the United Kingdom (in 2002, 2008 2011 and 2014), at least since the recent development of this nomenclature [39]. It is likely that this is partly due to the large number of isolates from the UK which have been sequenced and analysed. The UK isolates from the same sub-Clades as the endemic South African isolates of this study are likely travel related, imported to the UK from South Africa [39, 206]. There may be very closely related endemic populations in the UK, with a shared common ancestor and sub-Clade nomenclature to the South African endemic populations. These results highlight the links between the UK and South African *S. sonnei* populations as well as the ongoing role of international travel in *S. sonnei* epidemiology.

One likely endemic population also identified has likely not been identified before, being mis-identified as sub-Clade 2.8.2 by the sonneiTyping software. It is closely related to sub-Clades 3.7.7 and 3.7.9 and, therefore, to the other identified South African endemic *S. sonnei* populations. Many of the previously identified Clade 3.7 African isolates were from Northern or Central Africa, it is not unexpected to find a previously unidentified population [39]. The correct Clade identification but incorrect and incongruent sub-Clade identification, this being the only instance of incongruence between Clade and sub-Clade predictions, suggest that the scheme may need to be expanded for this under WGSA researched region to define nomenclature to this previously unidentified sub-Clade.

Previously identified African Clade 3.7 isolates belonged to sub-Clades 3.7.17 (Senegal and Morocco), 3.7.16 (Burkina Faso) and 3.7.12 (Kenya), suggesting that multiple sub-Clades of *S. sonnei* exist within Africa and are regionally associated [39]. The lack of overlap between the sub-Clades identified in this study and those previously identified for African isolates is therefore not unexpected. Indeed, the distance between the countries of origin for the previously characterised African *S. sonnei* isolates and

South Africa creates a hinderance to the effective *S. sonnei* population mixing required for these countries to have endemic strains in common.

As well as having a common geographic distribution across the country, the endemic strains have highly similar AMR profiles. Multidrug resistance is highly prevalent and typically included resistance to kanamycin, tetracycline, sulfisoxazole, streptomycin, streptothricin and trimethoprim. This agrees with previous phenotyping findings which has found MDR to be high in *S. sonnei* across the globe [194, 207-212]. Though the specific MDR profiles of populations are variable between region, many of the studied find high prevalence of resistance in *S. sonnei* to the antimicrobials identified in this study.

Resistance to these antimicrobials, in this study, was largely conferred by a single resistance gene across the endemic population, *tet(A)* for tetracycline resistance, *aph(6)-Id* for kanamycin, *dfrA1* for trimethoprim, *sat2* for streptothricin and *sul2* for sulfisoxazole. Against the exception, streptomycin, two highly prevalent resistance genes (*aadA1* and *aph(3'')-Ib*) were identified in the population, presence of both was common. Three of these genes (*dfrA1, tet(A)* and *sat2*) have previously been found in *S. sonnei* carried within a Tn7/in2 transposable element [35]. Meanwhile, three of the remaining genes (*sul2, aph(3'')-Ib* and *aph(6)-Id*) have been previously identified on a small plasmid (spA) [35]. Both mobile elements are associated with Clade 3.7, the presence of these genes in the study population is likely due to the presence of these mobile elements.

The virulence profiles are also highly similar across the endemic sub-populations, all having the iron metabolism *fec* loci and partial SHI-2 and enterobactin virulence loci (Table 5.5). It is unclear if the observed virulence profiles are typical of the global *S. sonnei* population or if the observed virulence profiles are region specific as *S. sonnei* virulence genotyping has not been widely performed. However, the sub-Clade 3.7.7 was associated with the presence of the pINV and several virulence genes (*capU*, *cea, traT, virF, ipaD*) which could be suggestive of differences in the virulence of this sub-population. However, as discussed further below, it seems likely that their presence is likely associated with the presence of the pINV rather than genuine differences in virulence gene prevalence.

Sub-Clade 3.7.11 was associated with recombination in a region encoding a known stress induced operon and part of a tryptophan synthesis operon. It is unclear what effect this recombination may have in the 3.7.11 sub-Clade isolates but suggests there may be stress response and metabolism differences in this sub-Clade compared to the other endemic strains.

5.4.1.2. *Likely imported strains*

As well as identifying several endemic strains, this study found evidence of imported *S. sonnei* strains. The imported strains include four more distantly related Lineage 3 isolates, belonging to Clade 3.6, which is associated with Central Asia, and Clade 3.4, which is associated with Latin America [39].

Both sub-Clade 3.6.1 isolates had resistance associated QRDR point mutations. The sub-Clade 3.6.1 isolate was phenotypically quinolone resistant, conferred by a single QRDR point mutation; while the sub-Clade 3.6.1.1.3 isolate was phenotypically fluoroquinolone susceptible but had triple QRDR mutations which should confer FQR [80]. The incongruency between QRDR mutations and phenotype might be due to human error during laboratory handling. Resistance conferring QRDR mutations are characteristic of the 3.6.1 sub-Clade [35, 39]. While the sub-Clade appears to have originated in Central Asia, it has since spread around the world [73, 78, 79].

Several rarer Lineages and sub-Clades were also identified among the South African isolates. One isolate belonged to Lineage 5, which has been predominantly found in Latin America but has also been detected in Europe and previously in South Africa [39, 72, 206, 211]. Meanwhile, a small cluster of sub-Clade 2.8.2 was also detected, one previous isolate of which was identified in Japan in 1964 [39]. The small number but close relatedness of the sub-Clade 2.8.2 isolates makes it unclear if these are imported strains or a local strain present at a very low level.

The South African 2.8.2 isolates were identified in Gauteng (2011 and 2013) and Eastern Cape (2013) provinces. These two provinces are a great distant apart but have two of the largest cities in South Africa and attract people from across the country and beyond. It would be possible for all these strains

to have been imported from the same region on multiple occasions, accounting for their close relatedness. Isolates from Tanzania and Madagascar have been previously identified as sub-Clade 2.5.1 [39]. It is possible that Lineage 2 is associated with Eastern Africa, possibly acting as a reservoir for strain importation of the Lineage.

5.4.1.3. *Strain emergence and pathogen population dynamics*

The time-based population structure is well supported in this study and the agreement of the dating estimates agree with the literature provides further support. Despite a narrow sampling time frame, the estimated most recent common ancestor (MRCA) dates from my Bayesian tree model line up with previous estimates: The MRCA of all the *S. sonnei* study isolates (median = 1786, 95% HPD = 1714-1849) overlaps with the previously estimated emergence time for the known global *S. sonnei* (median = 1669, 95% HPD = 1554-1763) [35]. Due to the detection of some widely diverse imported strains, a large portion of the genetic diversity of the known global phylogeny was captured within the data so a study isolates' MRCA dating estimate overlapping with the estimate for the MRCA of entire global *S. sonnei* population is to be expected.

The dating of the MRCA of all Lineage 3 study isolates (1967, 95% HPD = 1958-1978) also overlaps with the previous estimates (1972, 95% HPD = 1964-1979) [35]. Though the study estimate dating for the MRCA for the study Clade 3.7 (the Global lineage III) sub-populations was estimated to be more recent (median = 1993, 95% HPD = 1990-1996) compared to previous estimates (1972, 95% HPD = 1964-1979). However, the South African sub-Clades are closely related to each other, thus the MRCA of the South African Clade 3.7 isolates is not, therefore, the global Clade 3.7 MRCA but a younger descendant. An African / South American cluster within Clade 3.7 has been previously identified with a median estimated emergence in 1982 which falls within the 95% HPD estimates for the South African Clade 3.7 isolates MRCA from this study [35].

All the identified endemic sub-Clades likely emerged around the same time, 1996-1998 (95% HPD range: 1993-2001) the based on their respective MRCA. The coinciding of *S. sonnei* sub-population

emergence with a rapid increase in HIV prevalence in South Africa (throughout the 1990's with anti-retroviral treatment introduced in 2004) suggests that HIV infection likely played a role in *S. sonnei* epidemiology. A link between HIV infection and *Shigella* has been previously noted in the literature, which suggests a greater risk from shigellosis for those infected with HIV. It is likely, then, that the increase in population size and population diversification observed in this study were at least partially driven by the HIV epidemic in the country. If true, there may be a link between the HIV prevalence, historically, and the endemic *S. sonnei* population diversity observed in a region.

The literature suggests that *S. sonnei* populations are highly clonal outside of Europe due to the recent global spread of Clade 3.7 specifically [35, 73]. While several co-existing sub-Clades have been identified in South Africa, they are highly related and show very similar AMR and virulence profiles, likely the consequence of population expansion. Though the literature suggests that AMR acquisition is a common advantage aiding success in *S. sonnei* [35]. The similarity of the endemic sub-Clades likely means none has a competitive edge which could lead to clonal replacement, instead enabling co-existence.

The sharp increase in *S. sonnei* population size from around 2008 estimated in this study, is supported by the reported case numbers in the GERMS-SA annual reports which show that *S. sonnei* accounted for around 3% of shigellosis cases in 2007 but around 26% in 2008 [162, 163]. This is likely due to increasing industrialisation of South Africa [35, 161].

As previously noted, the study estimated South African endemic sub-Clades' MRCA overlapped with previous estimates for a MRCA of a cluster associated with Africa, suggest that the clonal replacement occurred on a multi-national scale within Africa. If true, the subsequent diversification of the new dominant strain, through genetic drift, would explain the presence of the multiple, highly similar, endemic sub-Clades observed in this study.

As previously mentioned, there is high similarity between the sub-Clades, offering little information of relative competitiveness. The emergence of the endemic sub-Clades at around the same time, however, means that the relative number of isolates per endemic sub-clade could be an indicator of their relative success of establishing in South Africa.

Sub-Clade 3.7.7 is the second smallest endemic sub-Clade yet is associated with the presence of the pINV, a large virulence plasmid required for enteroinvasive infection [201]. Absence of the pINV would likely render *Shigella* non-pathogenic; the lack of detection of the pINV is likely due to the plasmid being dropped during culturing than an *in vivo* absence [201]. The sub-Clade 3.7.7 association with the pINV suggests that the plasmid is maintained better outside of the host environment in this population, which likely has a fitness cost [24, 201]. Further study would be needed to prove that the 3.7.7. isolates have increased pINV retention, and that this retention has a fitness cost linked to *S. sonnei* success.

5.4.1.4. *Host population transmission dynamics*

The average age of infection was observed to be higher for girls and women compared to boys and men. I hypothesised that this was due to a greater risk of secondary infection in the home for adult women compared to adult men, and that this secondary infection likely arose due to adult women being more likely to be the primary care givers for children and sick relatives. Though the results of the relative transmission rate modelling were inconclusive, the median estimates did support my hypothesis of an increased risk of infection for adult women from children compared to adult men. The model is supported by the literature which also shows that shigellosis prevalence is highest in children under the age of five compared to other age groups as well as the GERMS-SA reported case numbers which showed average annual reported cases of *S. sonnei* attributable shigellosis to be 437 (319-567; 2011-2013), my model estimated a median population size of 331 per year (53-849) [10, 74, 150-152].

The high population size estimates for adult women appear, however, in conflict with the literature, which suggests that adults have a relatively low infection burden compared to children under the age of five. Though it is possible that the grouping of adult men and women together has led to an unrecognised high disease burden in adult women. If true, a high disease burden in women, would likely lead to transmission from adult women contributing to a large proportion of infections in men, as suggested by the model. Further work is necessary, however, to prove that secondary infection in the home contributes to a large portion of the risk of infection for adult women.

Previous modelling with age stratification and age and gender stratification of shigellosis in China found similar results to this study including high incidence in children under five years old and transmission from females to males being higher than the converse [213, 214]. Some differences were also found compared to this study such as the highest transmission rates occurring from adults (≥25 years old) to children under the age of five. It is likely that different cultural practices between China and South Africa account for some of these observed transmission differences, such as elderly relatives in China frequently looking after young children contributing to a higher transmission rate from the elderly to young children.

If my hypothesis of secondary infection in the home is correct, the stronger association between age of infection and gender with *S. sonnei* compared to *S. flexneri* may indicate that a higher proportion of *S. sonnei* transmissions occur directly from person to person compared to *S. flexneri,* which would likely have a greater level of transmission involving environmental passage such as through contaminated water.

Transmission of *S. sonnei* directly, person-to-person with minimal environmental passage, as the dominant transmission mechanism could explain the reduced success of a sub-clade which was better at maintaining the pINV outside the host. Active pINV maintenance would be redundant if the strain only exists within the host, where the environment promotes pINV presence, minimising any fitness benefit gained from active maintenance [24]. Any strain which stopped actively maintaining the pINV

would be able to minimise the fitness costs without losing the fitness benefits of having a pINV. In such conditions, maintenance of the pINV could be linked to success. Instability of the pINV in *S. sonnei* is known to be temperature dependent, with greater instability at environmental temperatures, providing further support for direct contact being the dominant transmission pathway for *S. sonnei* [24].

5.4.2. Antimicrobial resistance

This study confirms MDR is widespread in South African *S. sonnei,* with no observable differences in AMR profiles in regions across the country. Presence of low-level resistance to the important first and second-line treatments (fluoroquinolones and cephalosporins) also found in *S. sonnei*.

5.4.2.1. *Antimicrobial resistance of S. sonnei in a global context*

The AMR profiles of the endemic populations largely agreed with what has been found previously. The globally successful Clade 3.7 was previously identified as being strongly associated with MDR; 100% in the Clade 3.7 isolates appearing to possess *dfrA1*, *sat2* and *aadA1* AMR genes, conferring resistance to trimethoprim, streptothricin and streptomycin [35, 39]. Nearly all endemic isolates in this study did possess these genes, though occasional loss was observed. These genes have been previously identified as being carried on a transposable element so a lack of fixation in the population is not unexpected. The mobility of these genes suggests that given the right conditions these strains could become largely susceptible again. Prevalence of *aph(3'')-Ib, aph(6)-Id*, *sul2* and *tet(A)* AMR genes were also expected to be high, based on the AMR profiles of previously characterised African Clade 3.7 isolates [35, 39].

This study showed evidence of imported FQR and quinolone resistant strains from Asia, two of the *S. sonnei* isolates with resistance conferring QRDR mutations were part of Clade 3.6 which is associated with Central Asia [39]. The genetic distance between these isolates and the other study isolates, and the relatedness to strains from Asia suggest these are likely imported strains. The spread of quinolone and fluoroquinolone resistance from Asia fits with prior observations of spread to Europe, North

America and Australia [39, 78, 79]. There was little evidence to suggest that these strains were becoming established in the country, though a longer duration study would be needed to prove this either way.

Perhaps contrary to the literature, there was also evidence of multiple instances of *de novo* QRDR mutation occurring in the endemic strains. While none of these isolates had triple QRDR mutations, which is associated with FQR, it does show that decreased susceptibility may be present and that quinolone resistance is potentially evolving in the country. The emergence of quinolone resistant strains shows there is an evolutionary pressure for the development of quinolone, and possibly fluroquinolone, resistance in the country.

Previous research has pointed to Central Asia as reservoir for the emergence and dissemination of FQR strains, this study suggests that Central Asia may not be the only FQR reservoir [39, 78, 79]. The prevalence of QRDR mutations detected in this study, and the lack of *de novo* triple mutation, compared to the prevalence of FQR strains identified in some Asian countries it seems likely that the selection pressure for QRDR mutations is weaker in South Africa [78, 79]. Selection pressure likely varies across South Africa, however, as tuberculosis is associated with South African mining communities and fluoroquinolones are used to treat MDR/XDR TB (Iturriza-Gómara, personal communication, see acknowledgements above). Mining communities are mainly found in Free State, North West, KwaZulu-Natal and Mpumalanga provinces, which are largely rural provinces.

Local acquisition of cephalosporin resistance genes was also identified in *S. sonnei* though it is again unclear if this has influenced the success of these resistant strains due to the very low prevalence. The local acquisition of resistance to the currently recommended first and second-line treatments for treating shigellosis suggests that ongoing effective treatment is likely to become harder in the region. Showing the importance of defining AMR profiles and of developing an effective vaccine and alternative treatments.

5.4.2.2. *Antimicrobial resistance of S. sonnei in sub-Saharan Africa*

The endemic Clade 3.7 populations have highly similar AMR profiles with other, previously identified, African Clade 3.7 isolates, despite forming separate clusters within the Clade [39]. The similarity could point to similar selection pressures across the continent; however, most of the Clade have very similar AMR profiles. This Clade is found around the world, with international travel of the different sub-Clades likely occurring repeatedly. The high level of resistance of this Clade is believed to be the reason for its global success. The results of this study show that South Africa is dominated by this globally successful Clade and the prevalence of AMR in the country reflects this dominance.

Not all isolates from sub-Saharan Africa belong to the global Clade 3.7 [39]. Isolates from Madagascar and Tanzania were Lineage 2 isolates which are generally less likely to be MDR [35, 39]. It is possible that these countries are also dominated by Clade 3.7 as few *S. sonnei* isolates from these countries have been characterised through whole genome sequencing and subtyping using the recently developed nomenclature. However, the lack of Clade 3.7 isolates from this region so far could also be an indicator that Lineage 2 is dominant in Eastern Africa. Dominance by a different Lineage would likely mean a drastically different level of AMR to what was found in this study.

Phenotyping studies from across sub-Saharan Africa generally point towards high levels of MDR in *Shigella*. Though serogrouping of isolates is not always performed so the serotype specific AMR profiles of sub-Saharan Africa strains are hard to characterise from the literature. Where serotype has been reported though, MDR appears to be more associated with *S. flexneri* than *S. sonnei* [123, 124].

The resistance profiling of the endemic South Africa isolates appears to agree with phenotypic data from across sub-Saharan Africa. Resistance to amoxicillin has been observed to be high in other *Shigella* serogroups in sub-Saharan Africa but low in *S. sonnei* [118, 124]. The results from this study suggest that this would also be true in South Africa, as amoxicillin and ampicillin are both penicillin antibiotics and will likely be resisted by the same AMR determinants. This study found low AMR to ampicillin in *S. sonnei* but high AMR to ampicillin in *S. flexneri* 2a in the previous chapter.

This study also found resistance to both sulfisoxazole and trimethoprim be highly prevalent in both *S. sonnei* and *S. flexneri* in South Africa; in agreement with the literature from across sub-Saharan Africa which suggests resistance to cotrimoxazole, a combination treatment of trimethoprim and sulfisoxazole, is generally high in [114, 118, 123, 124].

The AMR findings in this study being as expected based on phenotypic data from across sub-Saharan Africa suggests that Clade 3.7 is probably the dominant Clade of *S. sonnei* across the continent. If true, then the epidemiology of *S. sonnei* elsewhere in the sub-continent would likely look similar to that observed in South Africa due to the high levels of similarity between the different sub-Clades of Clade 3.7. More work is needed to characterise the *S. sonnei* strains from the other African countries to be sure, and this will be examined further in the Chapter 6.

5.4.2.3. *Antimicrobial resistance of S. sonnei compared to S. flexneri*

The results of this study and the study from the previous chapter show that the AMR profiles of *S. sonnei* and *S. flexneri* are distinct, even within the same country. The acquisition of these current first- and second-line treatments appears to be happening faster in *S. sonnei* compared to *S. flexneri*. This may be due to *S. flexneri* may face greater fitness cost from AMR than *S. sonnei*. The link between AMR and *S. sonnei* success has been shown in the literature, while the *S. flexneri* study of the previous chapter shows that AMR is not always necessary for *S. flexneri* success. In South African strains, MDR *S. flexneri* were generally resistant to a greater number of antimicrobials compared to *S. sonnei*. Pan-susceptibility was also associated with *S. flexneri,* however.

5.4.3. Virulence

The presence of the large virulence plasmid, pINV, was not found in all isolates despite being an identifying feature of *Shigella* and carries all the required genes for the enteroinvasive pathogenesis. This is likely due to the plasmid being lost during clinical microbiology diagnostic processing, as a *Shigella* bacteria would be unlikely to be able to cause shigellosis without this plasmid [201].

The presence of the pINV was associated with the sub-Clade 3.7.7. isolates, as was the presence of several other virulence genes (*capU, cea, ipaD, traT* and *virF*). At least two of these genes are highly likely to be carried on the pINV, *ipaD* and *virF*, based on the literature and observations from this study [26, 215]. Both genes are important for the virulence functions of the pINV, *virF* is a transcriptional activator which promotes the expression of other virulence genes encoded by the pINV and some chromosomal genes in a temperature dependant manner [26]. While *ipaD* is both a T3SS structural protein and signalling protein which detects contact with host cells and induces effector protein secretion [215].

Nearly all *capU* encoding *S. sonnei* isolate contigs matched against *Shigella* pINV sequences during a BLASTn comparison against the nt database, suggesting that for at least some isolates this gene is also carried on the pINV. Previously identified as being within a virulence operon (encoding *virK*, *capU* and *shf*) on pAA2 plasmid and surrounded by insertion sequences previous work suggested *capU* was likely carried on a transposable element [198]. The correlation between the presence of *capU* and detection of the pINV suggests it may have been inserted into the pINV, though the lack of detection of the other virulence operon genes suggest that it may not have been inserted as part of the previously identified operon.

The likely encoding of *ipaD* and *capU* on the pINV explains the association between these genes and the pINV retaining sub-Clade 3.7.7. The other virulence genes, *cea* and *traT,* associated with this population may also be carried on the pINV. Though it is also possible that *cea* and *traT* are not carried on the pINV and their association with sub-Clade 3.7.7 is unrelated. Proving that of any of these genes are encoded on the pINV would require more complete genome assemblies, for example through long read sequencing, with the aim of completely assembling the pINV of each isolate. If genuinely associated with sub-Clade 3.7.7, *traT* likely enhances the host immune evasion as it likely encodes a complement resistance gene and has been shown to enhance resistance of *E. coli* to phagocytosis [195].

The least likely of the sub-Clade 3.7.7 associated virulence genes to be carried on the pINV is *cea,* a colicin encoding gene [216]. Colicins are involved in intra-bacterial competition [217]. This sub-Clade is also negatively associated with another gene involved in intra-bacterial competition, *celb* [216]. It is possible that there is redundancy between these genes, which would account their sub-Clade associations.

2.1.1.1. *Virulence of S. sonnei compared to S. flexneri 2a*

Distinct virulence profiles were observed between the two serotypes. The *Shigella* virulence loci SHI-1 and SHI-2, first identified in *S. flexneri* 2a*,* were complete and highly prevalent in the *S. flexneri 2a* but only partially present in *S. sonnei*.

The type 1 fimbriae encoding *fim* locus was also associated with *S. flexneri 2a*. This locus is associated with uropathogenic *E. coli,* promotes bacterial cell adhesion to host cells [179]. The absence of several of the structural *fim* genes in *S. sonnei* suggests that only *S. flexneri* 2a is expressing functional fimbriae. This also suggests that *S. flexneri* 2a is more virulent than *S. sonnei*.

The greater virulence of *S. flexneri* 2a over *S. sonnei* is supported by the association *S. flexneri* 2a and a systemic disease presentation. However, the association between urban areas and *S. sonnei*, compared to *S. flexneri* 2a, indicates that socio-economic factors affecting the host population may be the cause, rather than pathogen factors. The accessibility of healthcare in rural South Africa is poorer than in urban areas. It is possible then that those seeking medical care in rural areas are waiting longer before presenting at hospital, thereby increasing the probability of developing systemic infection.

The iron metabolism *fec* locus was absent in *S. flexneri* 2a but highly prevalent in *S. sonnei* [83]. Though the lack of *fec*-aided iron metabolism may have been counteracted by the presence of the complete enterobactin locus in *S. flexneri* 2a. The enterobactin locus, only partially present in *S. sonnei,* is involved in iron transport, aiding the binding of extracellular iron and transporting it into the bacterial

cell [182]. It is likely that despite having different virulence genes, the ability of these serogroups to metabolise host iron is similar.

5.4.4. Conclusions

The results of this study show that MDR *S. sonnei* strains circulate in South Africa, with a particular association with Gauteng province. Low level resistance to important antimicrobials is also present, acquired locally and introduced via strain importation.

There is little evidence for pathogen factors having shaped the epidemiology of *S. sonnei* in South Africa during the study period due to the high similarity of endemic strains. There is evidence of likely environmental influence on the epidemiology, however, including level of urbanisation, patient risk factors and transmission pathways, though further research is needed.

Chapter 6

Shigellosis across sub-Saharan Africa

Preface

This chapter brings together the isolates from the previous results chapters together with the previously characterised sample set from across sub-Saharan Africa. The methodology of the original collection of these isolates are detailed in "The Global Enteric Multicenter Study (GEMS) of Diarrheal Disease in Infants and Young Children in Developing Countries: Epidemiologic and Clinical Methods of the Case/Control Study" published in Clinical Infectious Diseases [129]. The subsequent whole genome sequence analysis of *Shigella* isolates collected during the GEMS study was published in the research article "Pathogenomic analyses of Shigella isolates inform factors limiting shigellosis prevention and control across LMICs" in Nature Microbiology [37]. The details of the work carried out as part of those studies can be found in the relevant papers. The contributions of my collaborators to this study, beyond the collection and sequencing of isolates as carried out in the previous studies, are detailed in the table below.

Neil Hall	Whole genome sequencing
Rebecca J. Bengtsson	Provided accession numbers for the phylogenetic reference isolates, having created known global population structure-representative, serogroup isolate lists, and smaller serogroup representative isolate lists. Also provided a working core-SNP alignment generation pipeline and bioinformatics support.

6.1. Introduction

The previous chapters in this thesis have examined the *Shigella* population within a single region and have therefore been unable to examine the level of strain transmission within sub-Saharan Africa nor the similarity among *Shigella* populations across the region. Two prior studies looking at the global populations of *S. flexneri* and *S. sonnei*, having included some African isolates, were the first to use WGSA to begin to examine these questions [33, 35]. However, the lack of available samples has resulted in poor resolution and limited understanding.

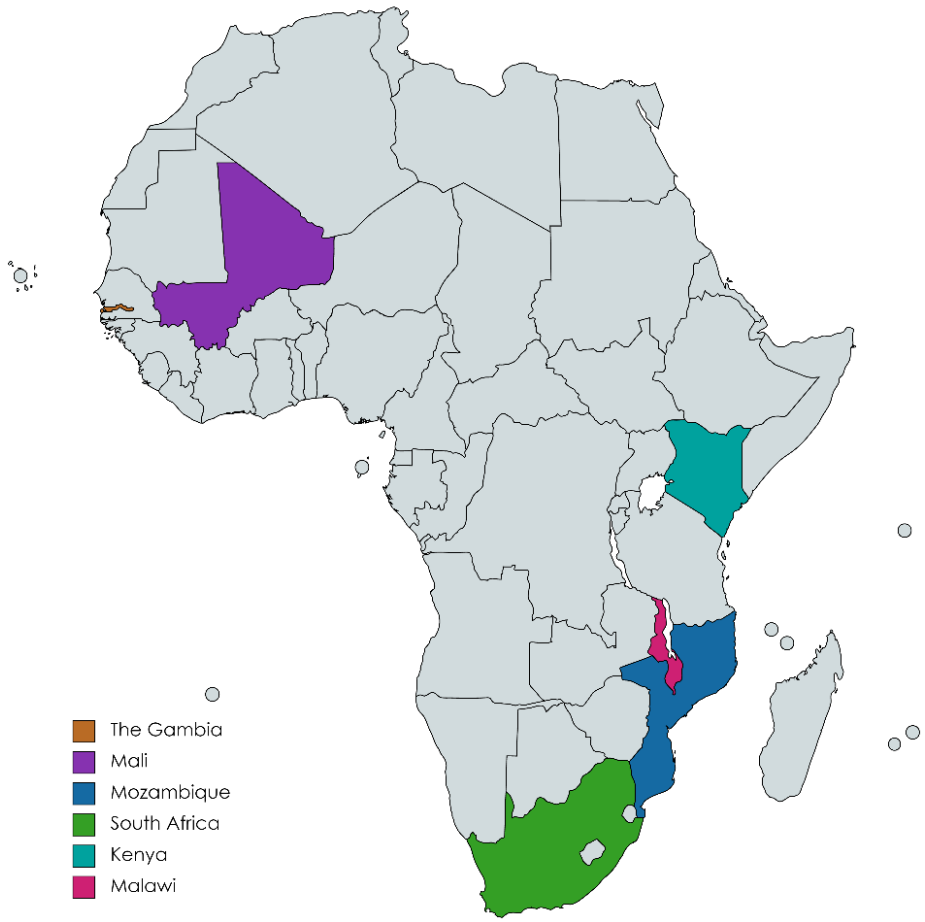

Figure 6.1. Sampled African countries.
GEMS study isolates collected in The Gambia, Mali, Mozambique and Kenya.

Whole genome sequence analysis on *Shigella* isolates collected as part of The Global Enteric Multicentre Study (GEMS), a large case-control study of enteric bacteria in children under the age of five with acute gastroenteritis across seven sites in Africa and Asia, has been conducted, but it has

not been used to specifically examine *Shigella* phylodynamics in sub-Saharan Africa [37]. Samples were collected for GEMS were collected across seven sites, four of which were African sites across four African countries (The Gambia, Mali, Mozambique and Kenya) (Figure 6.1. Sampled African countries. [14, 129].

In this study I aim to contextualise the newly described African isolates from this thesis within the previously described GEMS *Shigella* sample set, examining serogroup population structure and antimicrobial resistance profiles to better understand the relative similarity and potential level of strain transmission between sampled regions (Figure 6.1), and therefore across sub-Saharan Africa.

6.2. Methods

6.2.1. Sample selection

This study combines *Shigella* isolates from multiple datasets: 1) all *S. flexneri* isolates from the Malawian study described in Chapter 3, 2) all *S. flexneri* and *S. sonnei* isolates from the South African study, Chapters 4 and 5, and 3) *Shigella* isolates collected from African sites during the GEMS study, (Figure 6.1). From the newly described South African and Malawian datasets, only the isolates included previously, following quality control and *in silico* confirmation of serogroup, were included in this study; all previously excluded isolates were also excluded here.

Table 6.1. Breakdown of GEMS study isolates.
Showing both the number initially selected and then the final number of isolates included in the study. Serogroup of initial isolates is based on biochemical testing. Serogroup of final set isolates is based on in silico phylogenetic determination.

	Initial isolates (n)				**Final set** (n)			
	Mali	**The Gambia**	**Kenya**	**Mozambique**	**Mali**	**The Gambia**	**Kenya**	**Mozambique**
S. flexneri	25	79	63	29	22	74	60	28
S. sonnei	12	24	17	5	9	21	15	4
S. boydii	2	7	6	2	0	0	0	0
S. dysenteriae	1	5	19	0	0	0	0	0
Total	40	115	105	36	31	95	75	32

From the GEMS study, all biochemically identified *Shigella* isolates collected between 2007 and 2011 from children with acute gastroenteritis at African sites (The Gambia, Mali, Kenya, Mozambique) were initially selected for this study (296 isolates, all serotypes) (Table 6.1) [14, 129]. Isolates collected at Asian sites or from control subjects were excluded. The serogroup of all isolates was confirmed *in silico* with shigatyper (v1.0.6) and maximum likelihood phylogenetics (methods described below and in the Methods chapter) and any isolate demonstrated phylogenetically as not *S. flexneri* or *S. sonnei* were excluded from the study [202]. Following exclusion of isolates according to the above criteria, 184 *S. flexneri* and 49 *S. sonnei* GEMS isolates were selected for inclusion in the study.

6.2.2. Whole genome sequencing

All study isolates were subjected to whole genome sequencing. For isolates from previous studies, the quality trimmed sequenced reads from the previous study were again used for this study. For the Malawian isolates, this was performed according to in-house protocol at the Sanger institute (Chapter 3 section 3.2.1) [135]. The South African and GEMS isolates were both sequenced using at the Earlham Institute according to the Low Input Transposase Enabled (LITE) pipeline [136]. As with the South African isolates, some of the GEMS isolates (three *S. sonnei* and thirty *S. flexneri*) were re-sequenced due to poor sequencing quality, quality control methods are described below. Resequencing was performed at the Centre for Genomic Research (CGR, University of Liverpool). The sequencing was performed with the Illumina NovaSeq 6000 platform and DNA library was constructed using the NEBNext Ultra II FS DNA Library Prep Kit for Illumina [37]. Following re-sequencing, the new and original sequenced reads were merged prior to quality trimming and filtering.

Phylogenetic population structure reference isolates, detailed in Section 2.2, were included in the study, collected as part of prior studies and whole genome sequenced according to a range of methods [33-36].

6.2.3. Quality assessment and filtering

Quality assessment and trimming of isolate reads from the previous studies were described in the relevant chapters (Sections 2.3, 3.2.2, 4.2.1 and 5.2.1). Phylogenetic reference isolates were subjected to the same quality trimming and filtering as study isolates.

Sequence read quality was assessed using fastQC (v0.11.8) and multiQC (v1.7). All isolate reads were quality trimmed with trimmomatic (v0.38) and seqTK (v1.3) [55, 137, 138, 171]. Only isolate reads which met quality thresholds were included in the study, as described in detail in the Section 2.3. Two *S. flexneri* isolates were excluded from the GEMS isolates, one isolate because of a poor GC content curve, suggestive of contamination, one due to poor read mapping (<10x mean coverage across the reference *S. flexneri* 301 strain genome, chromosome: NC_004337.2 and plasmid: NC_004851.1).

Twenty of the *S. flexneri* isolates in the final dataset were re-sequenced isolates. No *S. sonnei* isolates from the GEMS sample set were excluded based on read quality.

Isolate serogroup was confirmed with serogroup specific maximum likelihood phylogenies (methods below in section 6.2.4) created using all *S. sonnei* or *S. flexneri* isolates. Included isolates were identified biochemically as *S. sonnei* or *S. flexneri,* or *in silico* predicted as *S. sonnei* or *S. flexneri* with shigatyper (v1.0.6). Isolates were required to be visually genetically related to all other isolates of the serogroup. Those which did not cluster within the serogroup phylogeny were removed from the phylogenies and new serogroup specific phylogenies were created.

Two biochemically identified *S. sonnei* isolates were excluded following *in silico* serogroup confirmation, while another was confirmed as *S. flexneri* and was therefore included with the *S. flexneri* isolates instead. Meanwhile, ten biochemically identified *S. flexneri* isolates were excluded due to not being *S. flexneri* or *S. sonnei*, with two were confirmed as *S. sonnei* and included with the *S. sonnei* isolates. One biochemically identified *S. dysenteriae* isolate was confirmed as *S. flexneri in silico* and was also included with the *S. flexneri* isolates. All *in silico* confirmed *S. dysenteriae and S. boydii* were excluded from the study.

6.2.4. Population structure and strain typing

All phylogenies were generated from core-SNP alignments using RAxML-ng (v0.6.0) [61, 62]. Core-SNP alignments were created from quality trimmed sequence reads and included phylogenetic reference isolates to provide the known global population structure.

Sequence reads were mapped to either the *S. flexneri* 2a 301 strain complete reference genome or *S. sonnei* 53G strain complete reference genome (chromosome: HE616528.1 and plasmids: HE616529.1, HE616530.1, HE616531.1, HE616532.1) using bwa mem (v 0.7.17). Variant sites were called, and a consensus sequence defined, using bcftools (v1.9), and core-SNPs were then defined using Gubbins (v2.3.4) with a filtering threshold set to exclude isolates with more than 31% missing data [144, 148,

218]. Isolates phylogenetically confirmed as not *S. flexneri* or *S. sonnei* were excluded and a new phylogeny created.

To examine the evolutionary relationships between the previously described South African and Malawian isolates and the GEMS isolates, single Phylogroup phylogenies were created where a mix of GEMS and Malawian or South African isolates was observed in a Phylogroup. Exclusively using study isolates, these Phylogroup specific phylogenies were created using the same methods as laid out above, with newly defined core-SNP alignments for just the relevant isolates using Gubbins.

Population clustering of the study isolates were created using RhierBAPS (v1.1.3) using isolate only core-SNP alignments and maximum-likelihood trees, generated in the same way as laid out above [219]. The population clusters defined by this software are referred to throughout as BAPS clusters.

6.2.5. Draft genome assembling

Draft genomes were assembled, with Unicycler (v0.4.7), for genotyping analyses [149]. Quality assessed in the same way as for previous studies, no isolates were excluded for poor assembly.

6.2.6. Antimicrobial resistance profiling

Genotypic and predicted phenotypic antimicrobial resistance profiles were generated for the GEMS isolates using the same methods as used in the previous chapters, as laid out in the Section 2.6 of the Methods chapter. AMR profiles for the South African and Malawian study isolates were also included for comparison.

6.3. Results

6.3.1. *Shigella flexneri*

6.3.1.1. *Population structure*

A mix of GEMS study isolates and the newly described Malawian or South African study isolates were identified in three Phylogroups, Phylogroup 2 (PG2), 3 (PG3) and the unnamed Phylogroup made up of *S. flexneri* serotype 6 (Sf6) (Figure 6.2). Only Malawian and GEMS isolates were found in PG2 and Sf6, while only South African and GEMS isolates were found in PG3. GEMS study isolates were also found in Phylogroups 1 and 6 (Figure 6.2). While one Malawian isolate (serotype 4av) was found in of Phylogroup 7 (Figure 6.2). As found previously, all South African isolates, being serotype 2a, were part of Phylogroup 3. Malawian isolate phylogroup distribution is as expected from serotype based on the literature [33].

To better compare newly described isolates with the previously described GEMS isolates, Phylogroup specific maximum likelihood trees were generated for those phylogroups where a mix of newly described isolates and GEMS study isolates were present (Figure 6.3).

Using RhierBAPS to define population clusters, based on genetic relatedness, I found three clusters in PG3 study isolates (Figure 6.3). The PG2 BAPS cluster 2 (BAPS2) contained only isolates from Kenya and Malawi while the PG2 BAPS3 cluster included isolates from across Africa, The Gambia, Mali, Kenya, Malawi (Figure 6.3). The two Malawian isolates in PG2 are closely related to isolates from both Kenya and The Gambia. While those from Mozambique, a neighbouring country to Malawi, were found in BAPS1 and so were relatively unrelated to those from Malawi, though the bootstrap support for this is not strong.

Like PG2, I found three clusters in Sf6, only one of which (Sf6 BAPS3) contained isolates from both Malawi and the GEMS study (Figure 6.3). This cluster was also exclusively Kenyan and Malawian isolates. All isolates from Mozambique belonged to Sf6 BAPS2, a cluster also containing isolates the Mali and The Gambia.

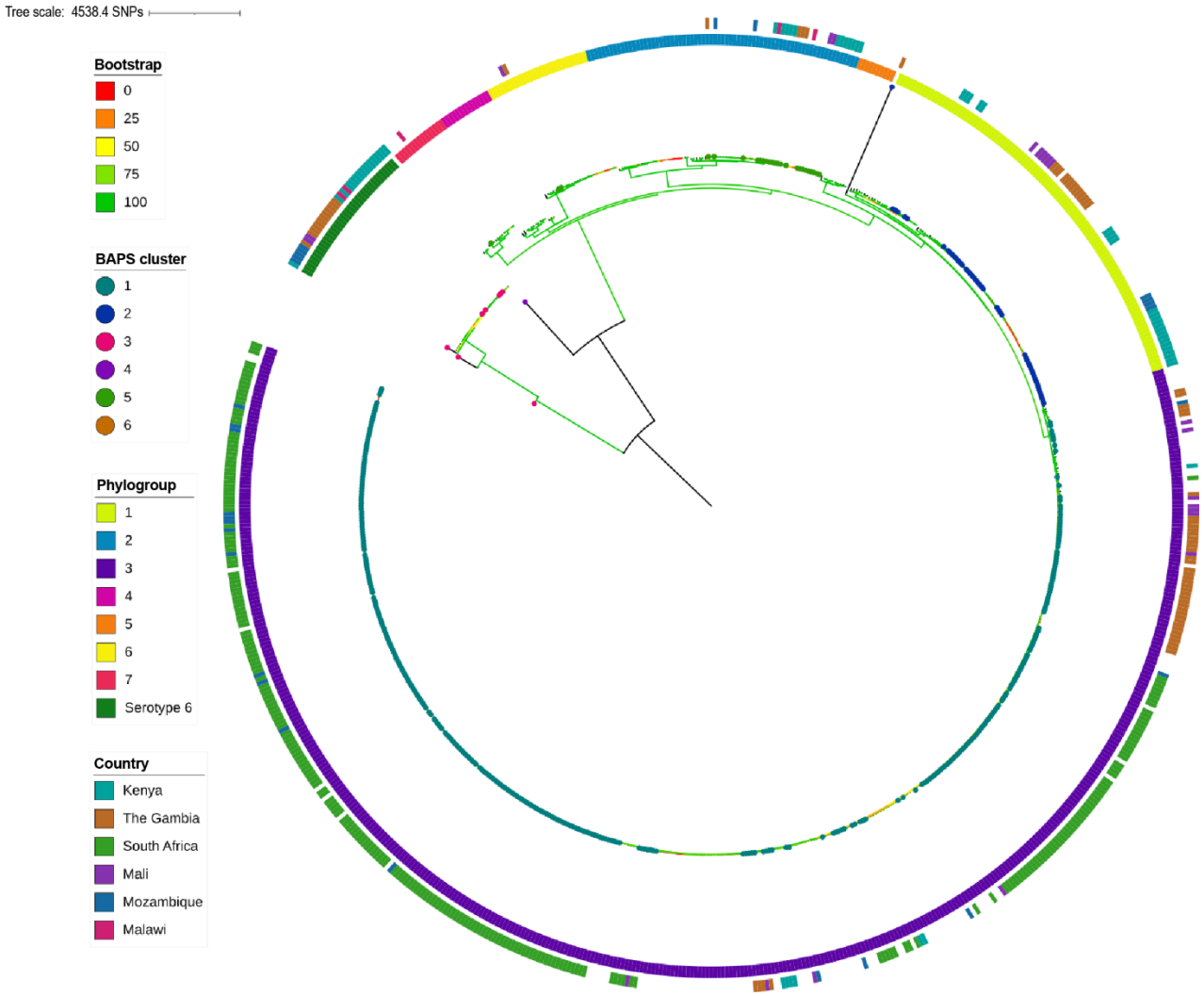

Figure 6.2. Mid-point rooted S. flexneri maximum likelihood phylogeny.
All phylogroups are included in the tree, indicated by bar colour in the inner circle surrounding the tree. BAPS cluster and country of origin are indicated for study isolates only. Terminal node colour indicates BAPS clusters while bar colour in the outer ring surrounding the tree shows country of origin. Bootstrap values are indicated by branch colour.

The PG3 tree shows that most GEMS isolates formed clusters (PG3 BAPS1 and 2) separately to the newly described endemic South African lineages (PG3 BAPS4, 5 and 6, and part of PG3 BAPS3) (Figure 6.3). The part of PG3 BAPS3 which does not form South African lineage 4 corresponds to a cluster of potentially imported strains identified in Chapter 4. The bootstrap values for the branches of the potentially imported strains are poor, however, so it is not possible to say for sure if these are imported strains nor where they may have come from (Figure 6.3).

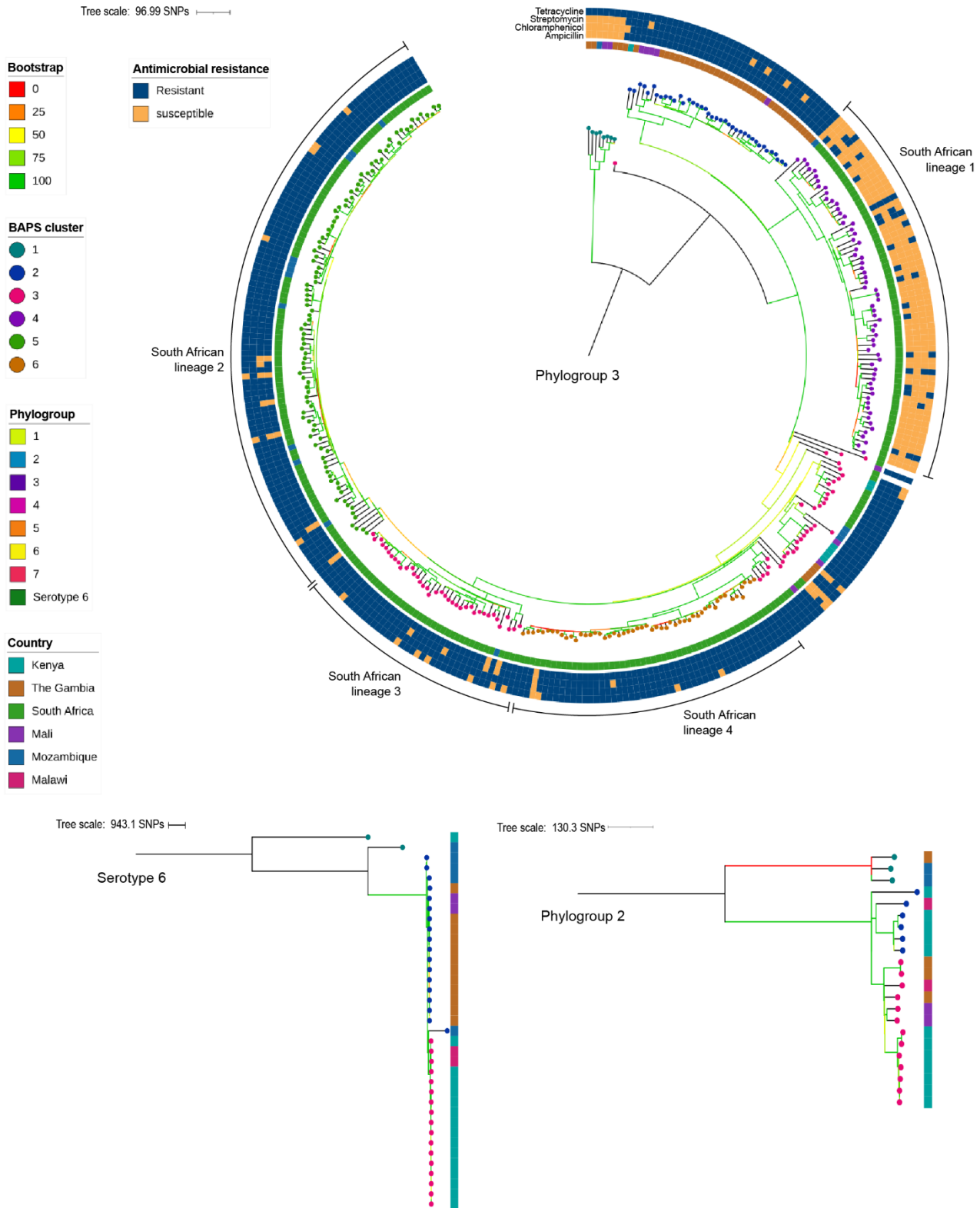

Figure 6.3. Mid-point rooted, phylogroup specific S. flexneri maximum likelihood phylogenies

Phylogroup 3 (top), Phylogroup 2 (bottom left) and the S. flexneri 6 Phylogroup (bottom right) all contain novel isolates from this thesis as well as GEMS study isolates. BAPS clusters were determined for each tree individually and are shown by terminal node colour. and country of origin are indicated for study isolates only. Inner ring bar colour shows country of origin. In the Phylogroup 3 tree the predicted resistance phenotype against four antimicrobials are indicated by the bar colour of the four outer rings. Bootstrap values are indicated by branch colour.

The results show that the GEMS isolates found within South African lineages 1, 2 and 3 were isolated in Mozambique, which borders South Africa (Figure 6.3). South African lineage 4 was exclusively found in South Africa (Figure 6.3).

6.3.1.2. *Antimicrobial Resistance*

The *S. flexneri* AMR profiles are highly similar in all the countries, apart from Malawi (Figure 6.3). Resistance to the same seven antimicrobials (ampicillin, chloramphenicol, kanamycin, streptomycin, sulfisoxazole, tetracycline and trimethoprim) was prevalent (≥55% of isolates) across all sites except Malawi where resistance to ampicillin and chloramphenicol was low (40%). All sites but Malawi have one dominant profile, one or two other prevalent profiles, and a range of low-level diverse profiles (Figure 6.3).

In Malawi, all isolates had a different AMR profile. Ciprofloxacin resistance, a fluoroquinolone, was also detected in Malawi, despite the small population size; the only site where FQR was detected apart from South Africa which has the largest sample size (Figure 6.3). This likely reflects the serotype diversity of isolates from Malawi.

Only one GEMS study isolate, from Mozambique, was found to belong to the *Shigella* resistance locus (SRL) lacking South African lineage 1 which is negatively associated with multidrug resistance (MDR) (Figure 6.2). This suggests that this sub-population is potentially specific to South Africa and that perhaps a successful pan-susceptibility associated sub-population is also specific to South Africa.

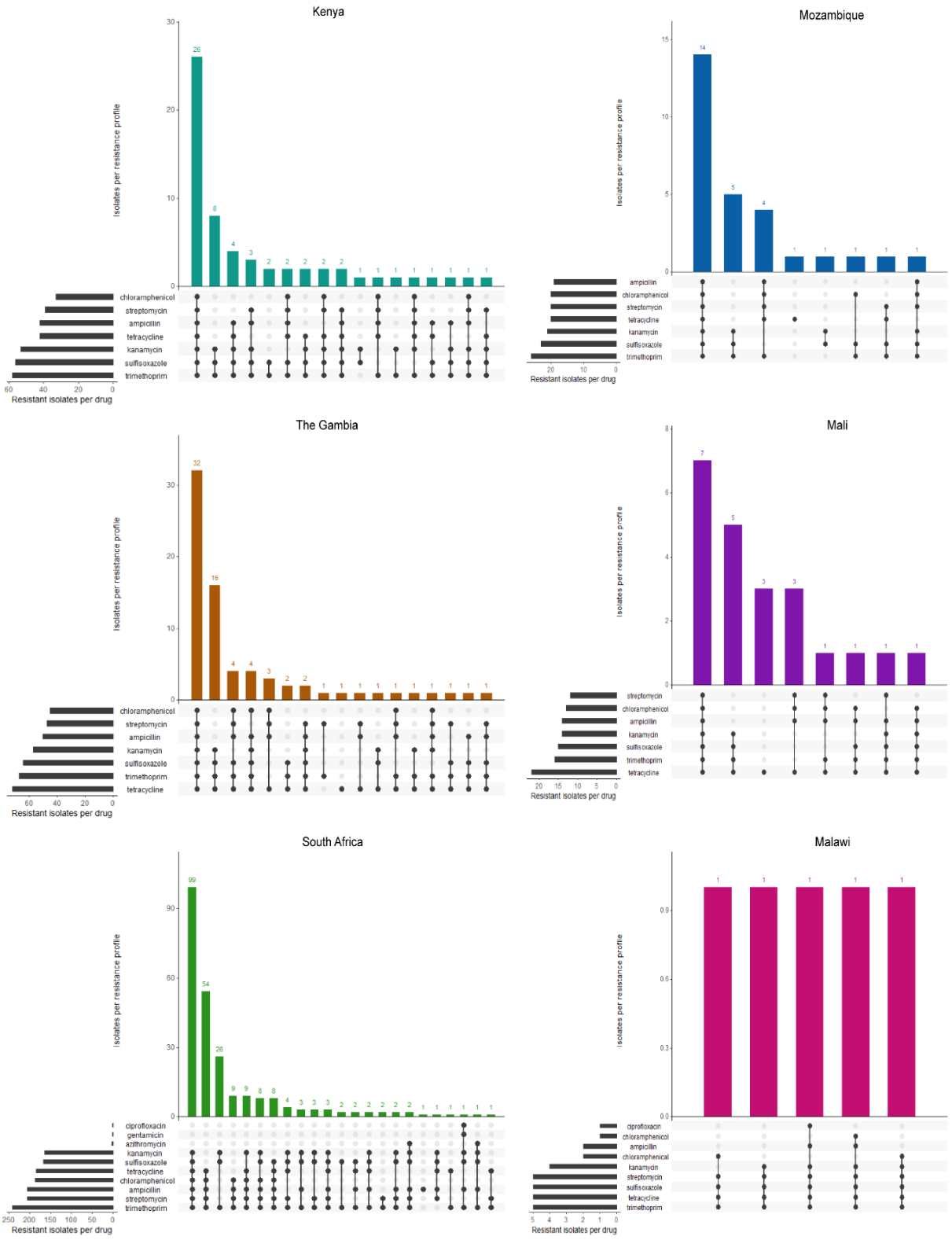

Figure 6.4. S. flexneri AMR profiles by country.

For each country, combinations of antimicrobials to which resistance was predicted (or tested where data available) are indicated by black dots to the right of the drug/determinant name, connected by a vertical black line. Vertical histogram along top of each figure part shows the number of isolates with each AMR combination (n above bar). Horizontal histogram to left of each figure part shows the number of isolates resistant to each antimicrobial. Kenya n = 60, Mozambique n = 28, The Gambia n = 74, Mali n = 22, South Africa n = 260, Malawi n = 5.

Another PG3 cluster (BAPS1) may also be lacking the SRL based on susceptibility (7/8 isolates) to three of the four antimicrobials (ampicillin, chloramphenicol and streptomycin) which the SRL encodes resistance to (Figure 6.2). Though this is suggestive of other PG sub-populations which lack the SRL across sub-Saharan Africa (The Gambia, Mali and Mozambique) this cluster is still associated with MDR, defined as resistance to three or more antimicrobial classes, carrying resistance genes to sulfisoxazole (7/8 isolates), tetracycline (8/8), trimethoprim (7/8), kanamycin (5/8).

Most of the other *S. flexneri* phylogroups appear to be associated with an AMR profile consistent with the presence of the SRL. However, based on GEMS isolates, the Sf6 Phylogroup also appears to lack the SRL; resistance to ampicillin, chloramphenicol and streptomycin was also low (3, 6 and 4 of 37 isolates respectively). Despite likely lacking the SRL, the Sf6 phylogroup was still associated with MDR (89% of isolates).

6.3.2. *Shigella sonnei*

6.3.2.1. *Population structure*

The mixing of the newly described South African and Malawi isolates with the previously described GEMS isolates were examined at the whole population level for *S. sonnei* as the genetic diversity of *S. sonnei* is far smaller than is observed in *S. flexneri* (6572 SNPs vs 9699 SNPs) [37]. All BAPS cluster MRCA branches had strong bootstrap support (>98%). Though some sub-clusters within BAPS clusters had poor bootstrap support, this had little effect on the results reported below.

Most isolates from both GEMS and South African datasets were in Lineage 3 Clade 3.7, with sub-Clades strongly associated with country of origin (Figure 6.4). Both sub-Clades 3.7.9 (BAPS4) and 3.7.11 (sub-cluster of BAPS2) were exclusively identified in South Africa, however, the South African associated sub-Clade 3.7.7 (sub-cluster of BAPS1) was also identified in Mozambique. Showing that there is some potential spread of strains between regions.

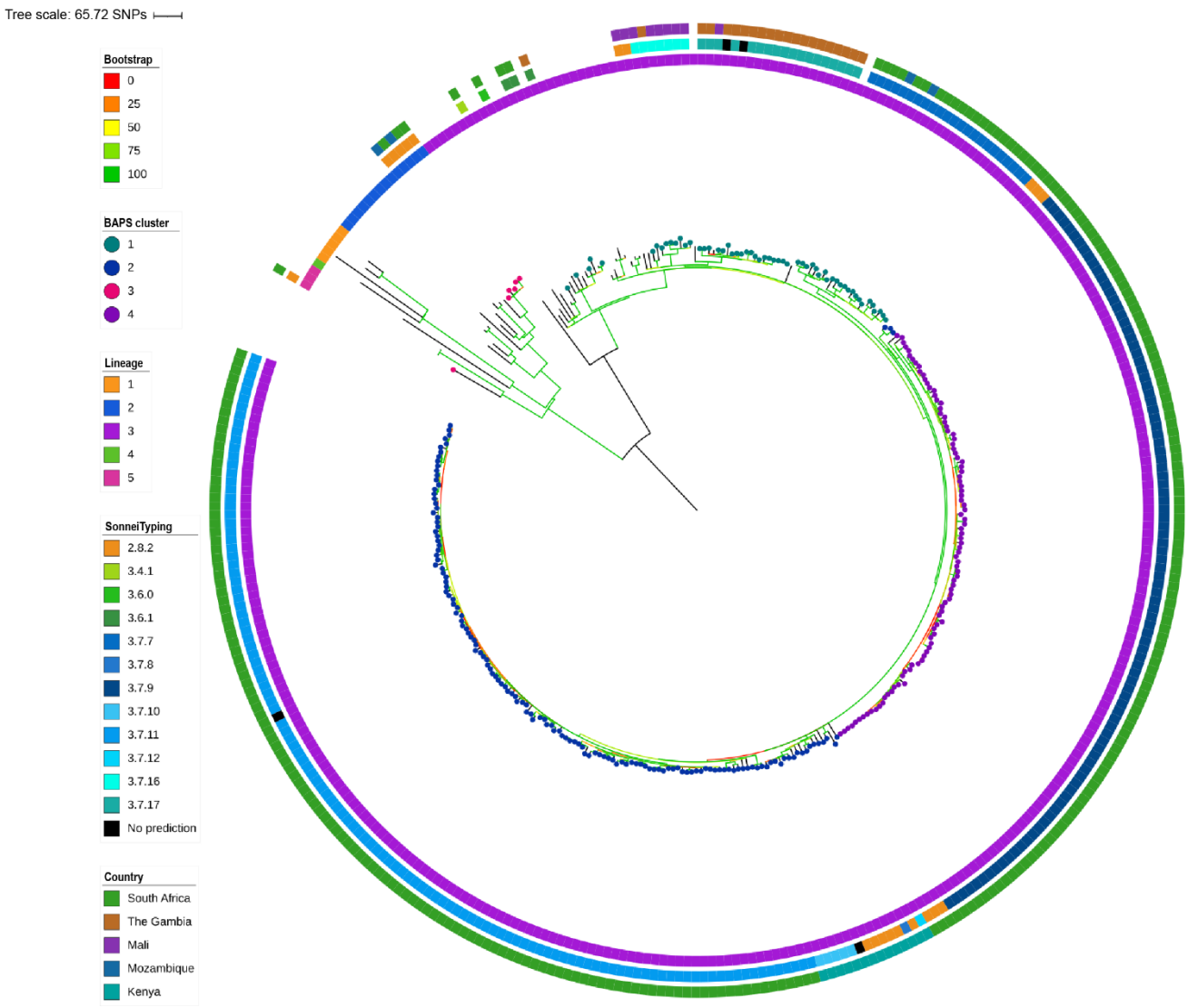

Figure 6.5. Mid-point rooted S. sonnei maximum likelihood phylogeny
All Lineages are included in the tree, indicated by branch colour. BAPS cluster, sonneiTyping phylotype prediction, and country of origin are indicated for study isolates only. Terminal node colour indicates BAPS clusters, inner bar colour shows sonneiTyping phylotype prediction and outer ring bar colour shows country of origin.

As was seen with the South African isolates, some of the GEMS Clade 3.7 isolates were mis-identified as belonging to sub-Clade 2.8.2, some of these were closely related to the Kenyan sub-Clade 3.7.10 (Figure 6.4). The two other 3.7 sub-Clades (3.7.16 and 3.7.17) were predominantly from Mali (89%) and The Gambia (95%), respectively, and were not identified in South Africa.

The cluster of Lineage 2 isolates, sub-Clade 2.8.2 (BAPS3 sub-cluster), was found in both South Africa and Mozambique (Figure 6.3); providing further support for transmission of *S. sonnei* strains within sub-Saharan Africa.

6.3.2.2. *Antimicrobial Resistance*

The *S. sonnei* populations AMR profiles were highly similar across all sites (Figure 6.5). A single dominant profile was observed with some low-level diversity at all sites except Mali; a greater level of diversity was observed in Mozambique than the other sites. Resistance to streptomycin, kanamycin, sulfisoxazole, tetracycline and trimethoprim high (>78%) at all sites except Mozambique (Figure 6.5).

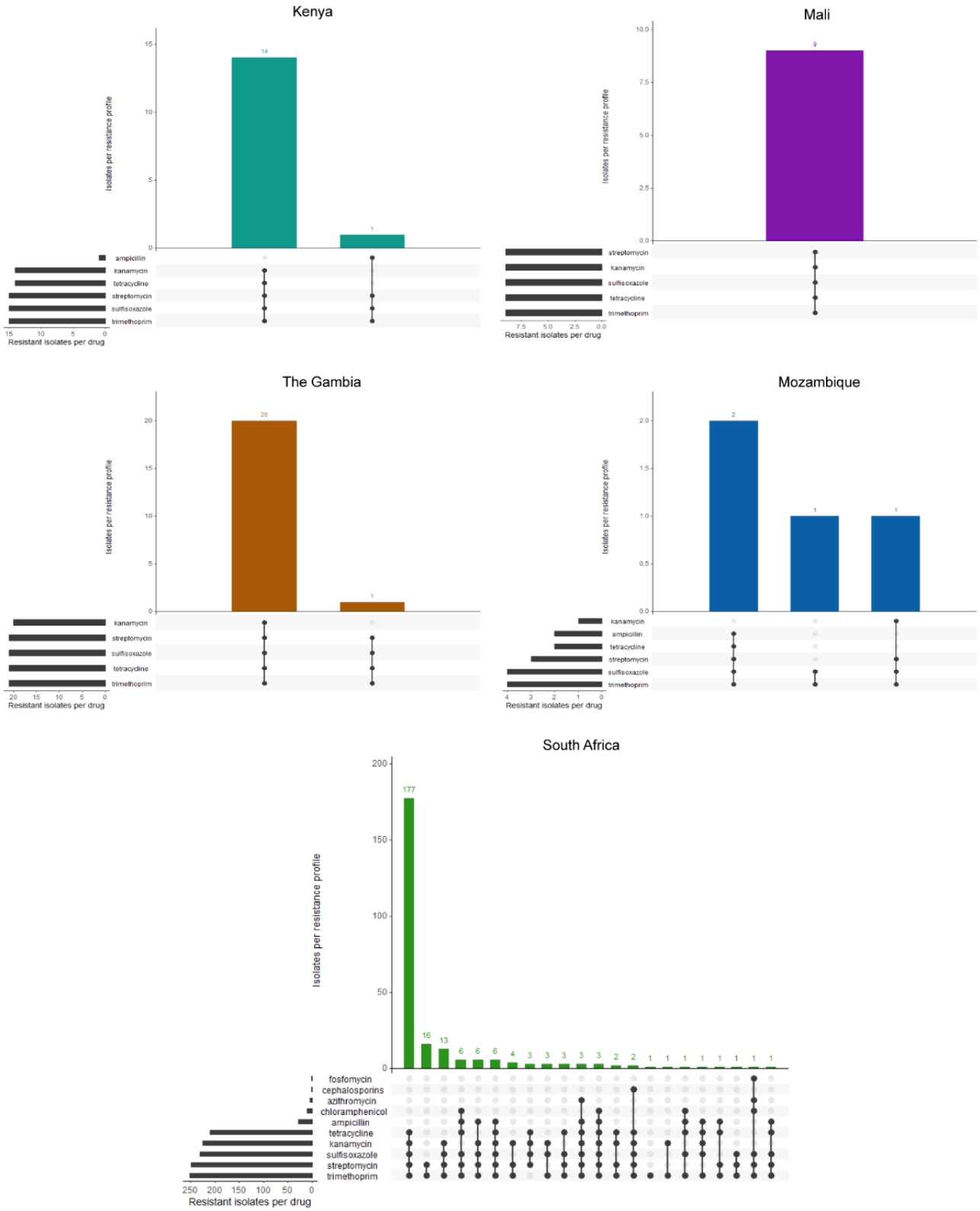

Figure 6.6. S. sonnei AMR profiles by country

For each country, combinations of antimicrobials to which resistance was predicted (or tested where data available) are indicated by black dots to the right of the drug/determinant name, connected by a vertical black line. Vertical histogram along top of each figure part shows the number of isolates with each AMR combination (n above bar). Horizontal histogram to left of each figure part shows the number of isolates resistant to each antimicrobial. Kenya n = 15, Mali n = 9, The Gambia n = 21, Mozambique n = 4, South Africa n = 253.

6.4. Discussion

This study shows distinct populations of *Shigella* by geographical region with some evidence of transmission between regions, predominantly between neighbouring countries, though cases of possible trans-continental strain transmission can be observed in *S. flexneri*.

The level of strain inter-regional spread appears to be greater in *S. flexneri* (12 possible instances to/from South Africa or Malawi) than *S. sonnei* (2 possible instances). It is likely that most inter-regional transmission region events occur via human migration, given the human host specificity of the pathogen, and thus is likely to occur at a similar frequency for both serogroups. The difference in levels of observed strain transmission between regions between the two serogroups may therefore be more reflective of strain colonisation success rather than the number of transmission events.

The greater spread of *S. flexneri* observed in this study could be due to a greater strain co-existence ability in *S. flexneri* compared to *S. sonnei.* Strain co-existence likely being possible due to relatively lower levels of inter-strain competition which would also improve chances for successful international transmission for *S. flexneri*. This would likely be cause by the reduced cross-protective immunity between *S. flexneri* serotypes. This is partially supported by similar levels of international spread between *S. sonnei* and most *S. flexneri* PG3 clusters which is predominantly serotype 2 [33]. It is, however, complicated by the diversity of likely inter-regional transmission events between sub-lineages of *S. flexneri* PG3 which suggests that not all sub-lineages are equally successful at transmitting between countries Figure 6.3. Mid-point rooted, phylogroup specific S. flexneri maximum likelihood phylogenies (Figure 6.3).

The high clonality of *S. sonnei* is likely a cause and consequence of ecological niche similarity between strains, resulting in high levels of inter-strain competition and which likely decreases the chances of successful international strain transmission, unless the imported strain has a competitive edge such as increased AMR [35, 72, 78, 79, 133, 206].

A greater proportion of isolates in this study were *S. flexneri* strains, confirming the dominance of the serogroup across sub-Saharan Africa [37, 95, 100-102, 104, 106, 118-122, 124-126]. The only region where the proportion of *S. flexneri* and *S. sonnei* were similar was South Africa and this was due to *S. flexneri* isolates being limited to a single serotype. When all serotypes are included, *S. flexneri* is also the dominant serogroup in South Africa [150-154, 162-164].

Evidence of endemic *S. sonnei* strains were found in all counties except Malawi, where no *S. sonnei* was detected, and perhaps Mozambique, where the low detection rate and genetic diversity of isolates could be indications of *S. sonnei* strain importation rather than endemicity. This is supported by the close relatedness of the Mozambiquan strains to South African strains. Strain importation into Mozambique would also explain high level of AMR diversity seen in the population compared to the *S. sonnei* populations in the other regions. A lack of endemic *S. sonnei* strains in Malawi and Mozambique it could be due to a low level of industrialisation of these countries [74].

Chapter 7

General discussion

7.1. *Shigella* evolution

7.1.1. HIV and diversifying evolution

7.1.1.1. *South African Shigella population diversification and the HIV epidemic*

Both South African studies found possible links between the HIV epidemic in the country and diversifying evolution in *Shigella*. The HIV epidemic was not unique to South Africa, however the country has had the largest HIV+ population globally since 2002, when it overtook India [187]. At this time the number of people living with HIV in South Africa (2.8 million) was greater than in North America, Europe, Latin America, and the Caribbean combined [187]. The size of the HIV positive population has since risen to an estimated 7.8 million people in 2020 (Figure 7.1) [187].

The size of the HIV positive population in the country makes South Africa both a unique and ideal place to examine the effects of HIV co-infection on the evolution of *Shigella*. It is likely that similar diversifying effects on *Shigella* evolution have, and are, happening across the world, to varying degrees, though this is the first time that such effects have been documented.

Evidence of diversifying evolution was found in identified endemic sub-populations and successful introductions of new endemic strains into the country (Chapters 4 and 5). The beginning of this diversification coincided with the early part of the HIV epidemic in the country (Figure 7.1).

A likely mechanism of HIV on *Shigella* evolution is through increasing susceptibility to shigellosis and thus increasing transmission and population size, as opposed to affecting the pathogen directly. This is supported by studies in the MSM community which has shown HIV to be a risk factor for shigellosis in the community, alongside other behavioural risk factors [188, 189]. The correlation between estimated increases in population sizes of both *S. sonnei* and *S. flexneri* 2a in South Africa (Chapters 4 and 5) with increasing in HIV prevalence also supports this hypothesis (Figure 7.1).

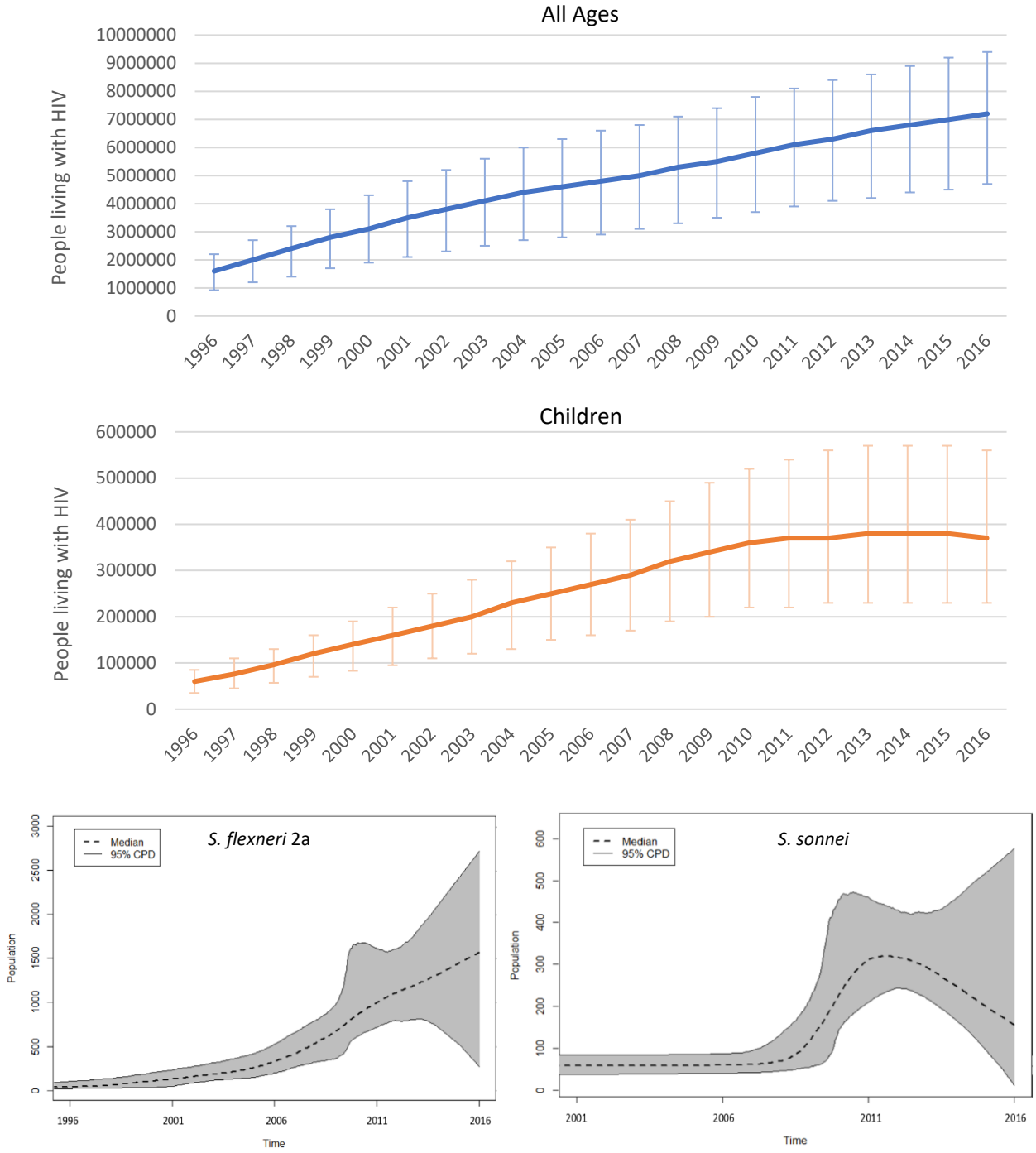

Figure 7.1. The number of people living with HIV in South Africa (top and middle), and South African Shigella population dynamics (bottom).

Number of people living with HIV in South Africa has been increasing since 1990, though the number of children under 15 years old (orange) peaked 2013-2014 and has since been decreasing. Estimates for South African Shigella population size suggests S. flexneri 2a population size has been increasing since 1996. Similar estimations for South African S. sonnei suggest population size increases only started 2006-2009 and may have started decreasing again around 2014.

HIV number estimates from the UNAIDS epidemiological estimates, 2021 [187]. Shigella population dynamics estimated during South African studies in Chapters 4 and 5.

There is a difference in the timing of population increase between the two serotypes (Figure 7.1). While there are several possible influencing factors for population dynamics, it is possible that effective HIV treatment roll out and post-Apartheid socioeconomic changes may have produced the observed differences.

The start of the HIV epidemic in the country also coincided with the end of Apartheid (1992-1994). Whilst Apartheid South Africa was industrialised, great efforts have been made post-Apartheid to reduce the legacy of socioeconomic inequality, from Apartheid, in the country. As dominance of *S. sonnei* in a country has been linked to the level of industrialisation it is possible the timing of post-Apartheid changes may have also influenced the timing of *S. sonnei* population dynamics (Figure 7.1) [74].

The end of Apartheid coinciding with the start of the HIV epidemic in South Africa adds complexity which obscures the relationship between the HIV epidemic and *Shigella* evolution in the country. It is likely that Apartheid ending also impacted shigellosis in the country, such as promoting the introducing of new *Shigella* strains into the country due to increased travel between South Africa and neighbouring countries, as discussed in Chapter 4.

7.1.1.2. *Impact of the HIV epidemic on the global Shigella population*

The lack of research, using WGSA, of shigellosis in sub-Saharan Africa means that there is little evidence from the literature to compare the results of this thesis with. Further study of *Shigella* in other sub-Saharan African countries with large HIV positive populations (Kenya, Mozambique, Nigeria, Tanzania, Uganda, Zambia and Zimbabwe), is needed to better understand the impact of HIV on *Shigella* evolution.

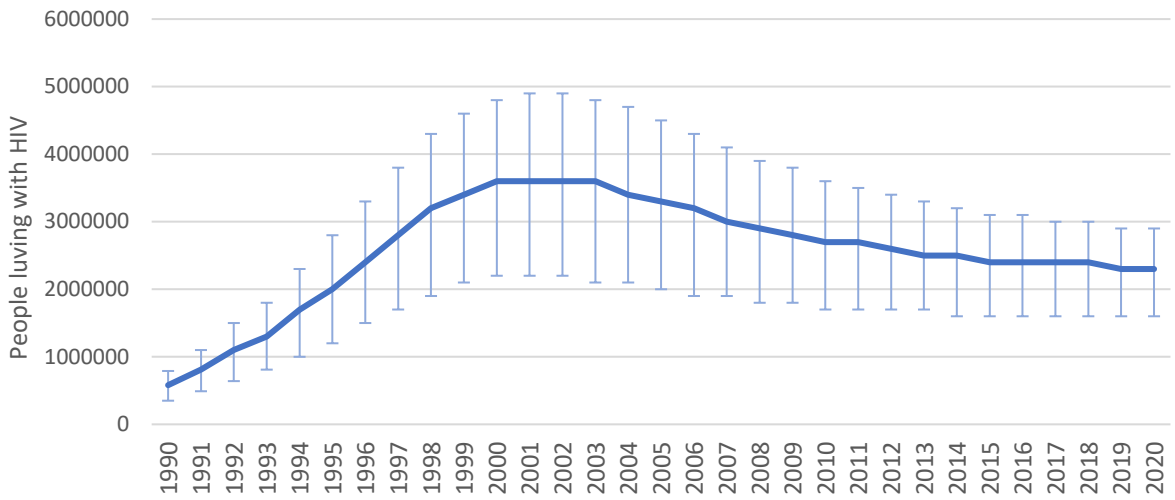

Figure 7.2. The number of people living with HIV, all ages, in India.
Number of people living with HIV in India peaked 2001-2002 and has since decreased. Estimates from the UNAIDS epidemiological estimates, 2021 [187].

Outside Southern and Eastern sub-Saharan Africa, India is one of the few countries which has also had a large HIV positive population (Figure 7.2) [187]. A small study of *Shigella* from patients attending the Department of Clinical Microbiology, Christian Medical College, Vellore, India with diarrhoea, between 1990 and 2017, estimated an *S. sonnei* population structure suggestive of serotype introduction to the region in 1995, followed by diversification until 2005, when an fluoroquinolone resistant (FQR) strain became dominant [222]. Multiple strains gained quinolone resistance associated point mutations in the QRDR between 2000 and 2005, including the FQR strain which became dominant post-2005, coinciding with the peak in number of HIV+ people living in the country (2001-2002 at 3.6 million people) (Figure 7.2) [187, 222].

Given the possible link between HIV and *Shigella* evolution identified in this thesis, it is possible that the accumulation of these point mutations may have been aided by the HIV epidemic. A link between HIV and FQR/quinolone resistance was previously identified in a study in the community of men who have sex with men (MSM), an association which was not seen for any other antimicrobial in the study [223]. Other studies have also failed to find a link between HIV and other AMR in *Shigella* [200, 223]. The exception was a study of shigellosis in Minnesotan, USA, residents which found HIV was associated with decreased susceptibility to azithromycin, both in MSM and with wider population [224].

Decreased susceptibility to azithromycin was higher in *S. flexneri,* however, which was found to be associated with HIV in two other studies [225, 226]. More study is needed to determine if HIV is associated with azithromycin resistance, or any other AMR, and to confirm the association with QRDR point mutations.

Most AMR in *Shigella* is acquired through AMR genes carried on mobile genetic elements, quinolone/FQR is, however, more typically acquired through point mutations [80]. Diversifying evolution is likely to increase the probability of a point mutation in the QRDR, and thus a (fluoro)quinolone resistant strain emerging, while having no effect on gene acquisition. This thesis found evidence of resistance-associated *de novo* point mutations in the QRDR in South African *Shigella* (Chapters 4 and 5).

Shigella strains endemic within the MSM community have been observed across the world [188, 227-230]. This has previously been attributed to the higher prevalence of risk-associated sexual behaviours in this community. Several studies have also found a link between HIV and shigellosis in men/the MSM community [188, 231-233]. The results of this thesis suggest that HIV is could be playing a role in driving shigellosis transmission, for both women and men, and the lack of prior evidence for an association between HIV and shigellosis in women is likely due the inclusion of too few HIV+ women.

7.1.1.3. *Linking HIV and resistance-associated point mutations*

Fluoroquinolone resistance is generally believed to have emerged predominantly in Southeast Asia [78, 234]. High use of fluoroquinolones in the region is certainly the dominant factor in emergence of FQR strains in the region, however the HIV epidemic may have also contributed.

A study looking at the emergence of FQR *S. sonnei* in Southeast Asia identified four separate emergences of new FQR strains, through *de novo* QRDR point mutations, once in 1995, once in 1996, once between 2001 and 2005 and once in 2007 [234]. It is not clear from the study in which countries these mutations occurred, and the HIV epidemic curve was different between the sampled countries

(Bhutan, Vietnam, Thailand, and Cambodia). Regardless of country, however, the HIV prevalence only ever reached the levels seen in the early stages of the HIV epidemic in South Africa, based on The World Bank total population sizes of each country [187, 235]. Rates of individual quinolone resistance conferring QRDR point mutations, and when they likely occurred, is not clear from the study [234]. Though the high level of FQR strain emergence supports dominant role of fluoroquinolone use for strain emergence.

Emergence of FQR *S. dysenteriae* occurred 1994 and 2002, in India and Bangladesh. Although the HIV prevalence in Bangladesh is unknown, the number of people living with HIV in India peaked in 2001 at 3.6 million people while in 1994 there was 1.7 million HIV+ people in the country [187]. The high numbers of HIV+ may have played a role in FQR emergence in India, in combination with the use of fluoroquinolones in the country.

A national study of *S. sonnei* in Vietnam observed multiple instances of *de novo* resistance-associated point mutations in the QRDR, though no FQR isolates were detected, following introduction of *S. sonnei* into the country around 1982 [73]. Subsequently spreading from Ho Chi Minh city to Khanh Hoa, circa 2006 and 2008, and to Hue, circa 2006. Population expansion typically follows a successful introduction and given the low prevalence of HIV in the country the observed expansion and emergence of QRDR point mutations is likely not linked to HIV in this case. This is supported by a recent study in which all Vietnamese isolates were FQR [236]. The subsequent emergence and dominance of FQR strains in the country is most likely due to fluoroquinolone use.

7.1.2. Pathogen ecology and niche specific evolution

7.1.1.4. *Variance in the stability of the large virulence plasmid and transmission route preference*

The South African chapters (Chapters 4 and 5) further confirmed the previously observed difference in pINV retention between *S. sonnei* and *S. flexneri*. The retention differences are due to differences in stability between the *S. flexneri* pINV (pINV$_{Sf}$) the *S. sonnei* pINV (pINV$_{Ss}$). The loss of the GmvAT and CcdAB toxin-antitoxin (TA) systems from pINV$_{Ss}$ makes the plasmid less stable than the pINV$_{Sf}$ which

has both of these TA systems [24]. It has been previously hypothesized that the reduced stability of the pINV$_{Ss}$ is a lifestyle adaptation, evidence of direct contact or host-to-host transmission with minimal environmental passage [24]. This is supported by the improved stability of the pINV$_{Ss}$ at human body temperature compared to environmental temperatures [24]. Active retention of the pINV at environmental temperatures enables *S. flexneri* to take advantage of transmission routes involving environmental passage.

This thesis finds, for the first time, a link between pINV stability and diminished success in *S. sonnei* strains, compared to strains with greater pINV instability (Chapter 5). Providing new evidence supporting the role of lifestyle adaptations in shaping *Shigella* epidemiology. Further research to better characterise the differences in the sub-population pINV$_{Ss}$ variants and other genome differences is needed, however, to confirm pINV stability as the main influencing factor in strain success in South Africa.

Further support for a link between high virulence plasmid stability and pathogen lifestyle can be found in other pathogenic *Enterobacteriaceae*. Both *Salmonella* and *Yersinia* maintain a large virulence plasmid and a lifestyle involving environmental passage.

Most virulence genes in *Salmonella* are encoded on the chromosome, however, one important virulence locus, the *Salmonella* virulence plasmid (spv) locus (*spvRABCD*), is plasmid-borne in *Salmonella enterica* subspecies I [237]. The complete spv locus encoding pSLT plasmid sequence shows the plasmid has two toxin-antitoxin systems, MvpAT and CcdAB [237]. The pINV$_{Sf}$ has MvpAT, GmvAT and CcdAB, while the pINV$_{Ss}$ has only MvpAT [237]. The stability of this plasmid is known to be high at human body temperature and, based on the presence of the two toxin-antitoxin systems, you would expect pSLT to be environmentally stable [238].

While host-adaptation, and thus epidemiology, can vary greatly between *S. enterica* Typhimurium strains, environmental passage is believed to be a part of all their lifestyles; both the host-specific *S.*

enterica Typhimurium LT2 strain and the broad-host *S. enterica* Typhimurium 14028 strain show long-term viability in soil [239, 240]. Both strains carry pSLT and have been shown to persist in soil for several weeks, successfully colonizing plants grown in the soil [239, 241].

Two species of *Yersinia*, *Yersinia enterocolitica* and *Yersinia pseudotuberculosis*, carry a required virulence plasmid, pYV. Both these species are leading agents of foodborne and zoonotic yersiniosis and likely have environmental passage as part of their lifestyle [237]. The pYV has been shown to be highly stable across a range of temperatures and pH conditions [242, 243]. At present, the specific genes and factors involved in the stability of the plasmid are unknown, though a ParDE family toxin–antitoxin system may be involved [237, 244].

A link between industrialisation of a country and *S. sonnei* dominance has been previously observed. Several possible reasons of this have been proposed, including pathogen lifestyle and reduced exposure to cross-protection providing *Plesiomonas shigelloidesis* [75]. Increasing industrialisation is typically associated with improved hygiene and sanitation infrastructure, leading to reduced exposure to raw sewage and contaminated water and thus reducing *Shigella* transmission via environmental passage. Given the differences in pINV stability, increasing industrialisation will likely favour transmission of *S. sonnei* over *S. flexneri*. Similarly, exposure to *P. shigelloidesis* likely reduces with increasing industrialisation; a bacterial species with an identical O-antigen to *S. sonnei* [31, 75]. Exposure to *P. shigelloides* provides cross-protection against *S. sonnei* but no other *Shigella* serotype.

This thesis provides support for lifestyle adaptation and transmission route preference between *S. sonnei* and *S. flexneri*. It does not, however, provide any evidence for or against the role of reduced cross-protection from *P. shigelloides* exposure with increasing industrialisation. The cross-protection from the shared O-antigen and adaptation to direct contact transmission are not mutually exclusive explanations for the success of *S. sonnei* in industrialised regions; it is likely that both mechanisms play a role in *Shigella* epidemiology.

The association between increasing industrialisation of a country and increasing dominance of *S. sonnei* strains supports pINV instability as an adaptation to a host-to-host lifestyle, a step towards becoming an obligate pathogen. Further evidence of adaptations towards being an obligate pathogen have also been identified previously, including the presence of a functional O-antigen capsule encoding operon, identified in *S. sonnei* but not *S. flexneri* [32]. The presence of the surface *S. sonnei* O-antigen capsule has been shown to modulate pathogenesis, reducing inflammation, and resisting complement-mediated killing, thereby increasing bacterial survival and persistence in the host environment [32].

7.1.1.5. *Antimicrobial resistance acquisition and finding success in a host-to-host lifestyle*

While acquisition of AMR determinants is strongly linked to success in *S. sonnei,* the link between AMR and *S. flexneri* success is more complicated (Chapter 4) [35, 73]. Acquisition of AMR determinants by *S. sonnei* typically leads to clonal replacement of the old strains by new more-resistant strain [35, 73]. Widespread multidrug resistance and dissemination of FQR strains suggests there are fitness benefits from AMR acquisition for is in *S. flexneri* [33, 78]. However, clonal replacement is not generally observed in *S. flexneri* as older strains persist and coexist with new strains [33]. With the identification of an emerging drug susceptible population in South Africa, this thesis proves, for the first time, that AMR is not always necessary for *S. flexneri* strain success.

The strong link between AMR and success of strains, regardless of serotype, in the MSM community, which is dominated by direct contact transmission, suggests that AMR is more strongly linked with success for pathogens with a host-to-host lifestyle [72, 133, 206, 227]. The lack of an observed association between HIV and AMR, aside from *de novo* point mutation-mediated (fluoro)quinolone resistance, in this thesis nor in the literature, suggests that HIV is not generally an influencing factor driving AMR acquisition in this community [200, 223].

Direct contact transmission is likely to dominate in places with high levels of access to clean water and sanitation provisions, thus I hypothesised that if acquisition of AMR was linked to a host-to-host

lifestyle, then MDR would be associated with urban regions of South Africa. No link between level of urbanisation and AMR was found (Chapter 4). However, level of urbanisation is not an accurate measure of industrialisation nor access to sanitation provisions in South Africa.

Informal urban settlements, known as townships in South Africa, are common and are typically underprovided for across many socioeconomic needs, including clean water and sanitation [245]. Informal dwellings are defined as "a makeshift structure not erected according to approved architectural plans" [245]. Settlements are grouped into four categories, 1) formal urban, 2) formal rural, 3) informal urban, 4) traditional [245]. These townships are often associated with a formal urban settlement, consequently places with the greatest proportion of people living in informal settlements, in 2020, were cities: Buffalo City, Eastern Cape (27.2% of people living in the city); Johannesburg, Gauteng (19.0%) and Cape Town, Eastern Cape (18.6%) [246].

7.2. An updated understanding of *Shigella* epidemiology

Beyond describing the genomic epidemiology of shigellosis in several regions across sub-Saharan Africa, this thesis has provided evidence for several factors influencing *Shigella* evolution and epidemiology, discussed above. The examined factors are interacting and overlapping in their influence on shigellosis epidemiology (Figure 7.3). Using the evidence from this thesis provides insights into the epidemiology of shigellosis across sub-Saharan Africa, and more broadly.

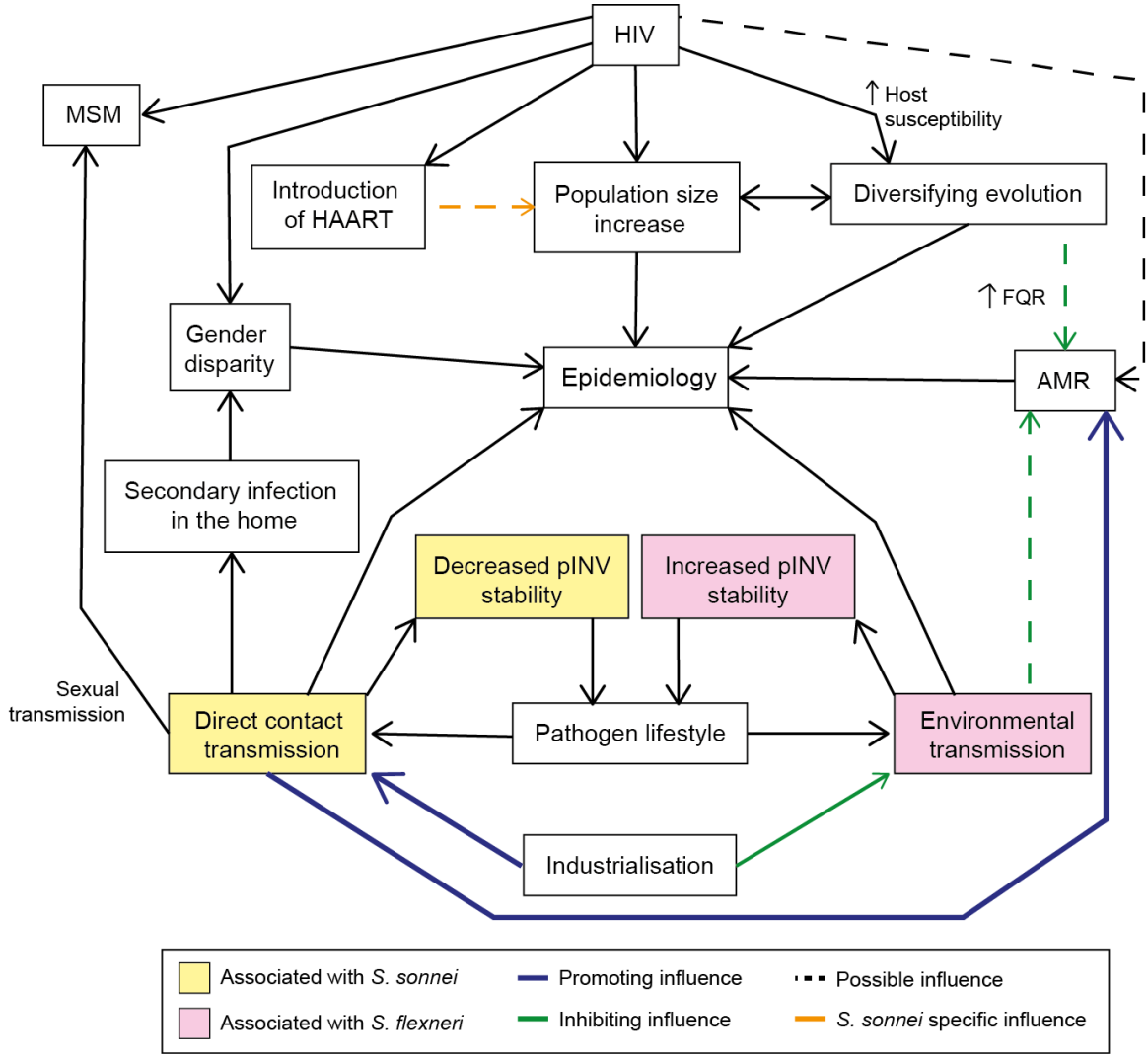

Figure 7.3. The interaction of multiple factors identified and described in this thesis which affect shigellosis epidemiology.

Pathogen lifestyle, hypothesised to be differently preferred by *Shigella* serotype, affects epidemiology both directly and indirectly through pathogen adaptations. Factors associated with *S. sonnei* are indicted by yellow boxes, and those associated with *S. flexneri* 2a by pink boxes. Solid black arrows show where a highly probable link between factors exists, dashed where a possible link exists, a blue arrow where a promoting influence on a factor exists and green where an inhibiting influence exists.

Strong geographical clustering (Chapter 5) and identification of at least one previously unidentified strain (Chapter 4) suggests that a large portion of the *Shigella* population remains unsampled across sub-Saharan Africa. With a high disease burden in the region this also means a large portion of the global *Shigella* population is also largely unsampled.

Several South African endemic strains were identified in this thesis. However, it was not possible to determine if these strains were introduced to the country from other sub-Saharan African countries, or if they were imported from outside Africa. Previous global population studies show that multiple successful introductions to the sub-continent have happened for *Shigella* [33, 35]. Multiple introductions of other enteric pathogens to sub-Sharan Africa have also been observed [247, 248]. Meanwhile, the results from Chapters 4, 5 and 6 show strain transmission within sub-Saharan Africa occurring.

Prevalence of HIV across sub-Saharan Africa is the highest in the world, particularly East and Southern Africa. The greatest probable impact of the HIV pandemic on *Shigella* will, therefore, likely be in sub-Saharan Africa, meaning that *Shigella* diversity in sub-Saharan Africa will likely be greater than has been generally observed in other regions. Comparing the diversity of strains between studies is difficult, however, due to the variety of study designs, methods used and the focus of reported results.

7.2.1.1. *Shigella flexneri*

In South Africa, it is possible that the epidemiology of serotype 2a is typical for *S. flexneri* serotypes in the country. Serotypes 1a, 3a and 6 which are known to be present, though less prevalent, in the country [150-154, 162-164]. If true, the other *S. flexneri* serotypes present in the country would also have multiple endemic lineages with distinct geographic distributions. Though the genetic diversity within these endemic lineages may be less than seen for serotype 2a, and there may be fewer endemic lineages, due to the smaller population sizes.

The high level of diversity of strains within serotype 2a, each with a distinct distribution, observed in South Africa is likely typical for *S. flexneri* across sub-Saharan Africa. The high level of diversity is supported by prior work showing that high genetic diversity is typical for *S. flexneri* serotypes [33].

Other serotypes are dominant in different sub-Saharan African countries, as discussed in the introduction (Chapter 1, Table 1.2). Highly populated countries in sub-Sharan Africa likely to have multiple strains of the dominant *S. flexneri* serotype, and probably even for the less prevalent serotypes. The results from Chapter 4 suggest that each strain may have distinct distributions within the country. While the results from Chapter 6 show that endemic strains are likely to be unique to each country.

Smaller countries will likely have less diversity owing to smaller *Shigella* population sizes. The evidence from Malawi (Chapter 3) and the GEMS study suggest that this does not necessarily translate to fewer serotypes present in the country [37]. The genetic diversity of serotypes is likely exacerbated by the high HIV prevalence across sub-Saharan Africa.

7.2.1.2. *Shigella sonnei*

The evidence from this thesis and the literature suggest that each sub-Saharan African country likely has a single *S. sonnei* strain, which is geographically associated with that country and maybe the neighbouring countries. Although, countries with very high levels of human travel in, out and around the country, or smaller countries with strong traffic links to neighbouring countries, may have a shared *S. sonnei* strain. Each strain may have several identifiable sub-populations; however, the similarity of these sub-populations is likely to be high enough for these sub-populations to be considered a single strain (Chapter 5). I expect that the introduction of new strains would only be successful if the new strain had a competitive edge resulting in clonal replacement [35].

Chapter 5 shows fewer observations of possible international strain transmission for *S. sonnei* than was seen for *S. flexneri* in Chapter 4. Fewer probable strain introductions could be due to lower

frequency of *S. sonnei* strain transmission than *S. flexneri* or it could be that successful onwards transmission is less common in *S. sonnei* and thus introduced strains are detected less often. Transportation of strains across national borders is likely to predominantly occur through infected host movements, rather than via contaminated food or water, regardless of serogroup. Importation through host travel is supported by the high proportion of patient travel-related cases in high income countries compared imported contaminated food-related cases [206, 212, 249-259]. Though more research is needed to show that there is no difference in intercontinental strains transmission between serogroups. The smaller number of possible intracontinental transmission are, therefore, most likely further evidence of clonal replacement rather than strain coexistence in *S. sonnei,* globally and across sub-Saharan Africa.

Studies which included serotype testing of *Shigella* isolates from sub-Saharan Africa generally identify *S. sonnei* less frequently than *S. flexneri* (Table 1.2) [95, 100-102, 104, 106, 118-126]. Less frequent detection suggests *S. sonnei* is not a dominant serotype across much of the region. This is supported by the results of Chapter 2, in which no *S. sonnei* isolates were detected in Blantyre, Malawi, as well as the literature which shows that *S. flexneri* is dominant in LMIC while *S.* sonnei dominance is associated with industrialisation of a country [74].

Not being the dominant serotype in a country likely means that the genetic diversity of *S. sonnei* strains in other sub-Saharan African nations will be less than was found in South Africa (Chapter 4). The *S. sonnei* populations will likely be smaller and have a less complicated population structure than for *S. flexneri* serotypes. This is supported by the genetic diversity of the likely endemic *S. sonnei* populations in the multi-national Chapter 5 study compared to the genetic diversity seen in the endemic *S. flexneri* 2a populations. Whole genome sequence analysis of *Shigella* isolates collected from multiple African and Asian sites during the GEMS study shows *S. sonnei* generally has far lesser genetic diversity than *S. flexneri* serogroups [37]. However, the diversity was highly variable between *S. flexneri* serotypes and *S. sonnei* diversity was similar to *S. flexneri* 1b and 2b.

7.2.1.3. *Antimicrobial resistance*

Antimicrobial resistance is a growing concern in Shigella. This thesis found widespread MDR across all serotypes and locations. Resistance against the few remaining widely effective antimicrobials was low, however.

The only population which was not found to be generally MDR was an emergent South Africa *S. flexneri* sub-population which lacked the SRL multidrug resistance element. No evidence of similar drug susceptible populations was identified in any of the other studies in this thesis. This may be because no other susceptible population exists. however, as distinct, sub-national distributions were observed for *S. flexneri,* and the other studies included samples collected at a single site per country, it is possible that susceptible sub-populations of the study countries were not detected. Similarly, many sub-Saharan African countries were not sampled from. The emergence of a drug susceptible *Salmonella* lineage in Malawi supports the probability of conditions existing in other sub-Saharan African countries in which a drug susceptible lineage may emerge [186].

7.3. Shigellosis and global public health

Several important factors which may impact public health policy decisions have been identified during this work. These factors include the impact of HIV coinfection, MDR, and pathogen lifestyle on transmission and patient demographic.

HIV coinfection has been previously linked to an increased risk of systemic disease and death from shigellosis [200, 220, 221]. Possible diversifying effects of the HIV epidemic in South Africa could be due to increased transmission and an increased *Shigella* population size (Chapters 4 and 5). Increased transmission could mediated through increased susceptibility to infection or chronic infection and long-term shedding. Thus, HIV may be a risk factor for shigellosis infection and/or coinfection maybe a risk factor for reduced infection clearance leading to chronic carriage. Data from a South African study found HIV to be a risk factor for systemic disease in adults but not children, suggesting that the influence of HIV-coinfection on shigellosis may be multifaceted and more research is needed [220].

While more research is needed to better understand how HIV affects shigellosis transmission, improving HIV control will likely aid in shigellosis control. Other interventions such as targeted risk-awareness and risk-reduction, education for HIV+ individuals, and post-treatment testing for HIV+ individuals to ensure infection clearance will also help reduce shigellosis. Offering HIV testing to patients following a shigellosis diagnosis may also aid in the reduction of HIV and the minimisation of the impact of HIV coinfection on shigellosis.

Older girls and women are at greater risk from shigellosis than their male counterparts in South Africa (Chapter 4 and 5) [200]. The differences in risk are likely due to the combined impacts of greater exposure, from looking after children and sick relatives, and the higher proportion of HIV seropositivity (Figure 7.3). Secondary infection in the home has been shown during a recurrent *S. sonnei* outbreak in New York, which was driven by young children attending school outside the home [260]. Support for the role of secondary infection is the greater difference in average age of infection between men

and women in *S.* sonnei attributable shigellosis compared to *S. flexneri* (Chapters 4 and 5), secondary infections in the home likely to involve minimal environmental passage.

HIV is higher in women compared to men for sub-Saharan Africa as a whole, both in terms of numbers of newly infected individuals in 2020 and the number of people living with HIV [187]. This means that women across sub-Saharan Africa likely face greater exposure to shigellosis whilst also being at greater risk of becoming infected and/or passing on the infection due to HIV. Public health policies targeting shigellosis in those older than fifteen years old should have a greater focus on women and girls to aid better management of these risks.

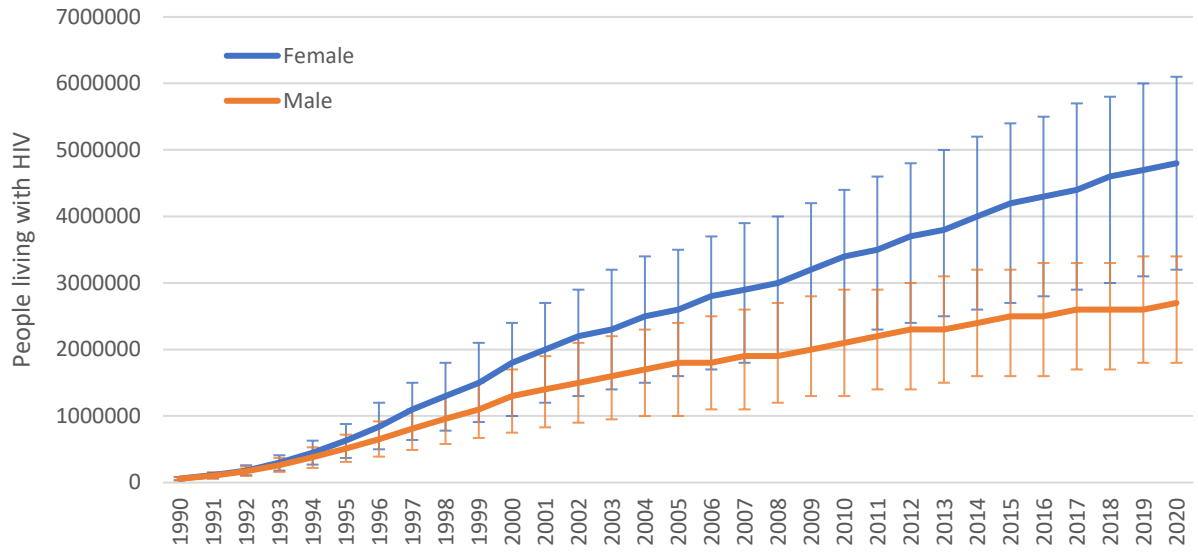

Figure 7.4. The number of people over 14 years of age living with HIV in South Africa, by gender.
The number of people living with HIV is higher in women than men. HIV number estimates from the UNAIDS epidemiological estimates, 2021 [187].

Differences in pathogen lifestyle suggests that different measures are needed to effectively reduce transmission. Reducing transmission via environmental passage, through improved sanitation and access to clean water, as well as hygiene education and regulations are food production, will likely have a greater impact on *S. flexneri* transmission than on *S. sonnei*. While hygiene education and targeted public health messaging will likely have a greater impact on *S. sonnei* transmission rates than on *S. flexneri*.

7.4. Context and future directions

7.4.1. Scope and limitations

The scope of this thesis project is affected by the study design in several ways, the main being the sample selection. In all studies, isolates were collected from patients attending a hospital for diarrhoea, though it is known that asymptomatic carriage of *Shigella* does occur [129]. The sampling from only 'case' isolates means that a portion of the *Shigella* population in the geographic study region is not represented in each study. Phylogenetic population studies of isolates from the GEMS case-control study, however, suggest that case and control isolates are highly mixed in the population [37]; symptomatic infection vs asymptomatic is likely more influenced by human host factors than pathogen factors. This suggests then that sampling only from 'cases' will have minimal impact on the range of *Shigella* population sampling and therefore on the scope of this project.

All isolates included in this study were collected at public hospitals. Sampling from only the public healthcare sector will reduce the scope of this project as public healthcare is typically accessed by a different demographic than private healthcare. The demographic of people accessing private healthcare will likely have better access to hygiene and sanitation provisions and will thus likely contract shigellosis via direct contact in a higher proportion than those accessing public healthcare. It is possible, therefore, that specific lineages are being transmitted via direct contact within the demographic of people accessing private healthcare, as with the MSM-associated lineages identified around the world [133, 261-263]. Sampling only from public healthcare would likely fail to detect these lineages, if they exist. However, no research directly comparing *Shigella* strains between public and private healthcare has yet been conducted, to my knowledge.

Uneven sampling is further exacerbated by the variable access healthcare by geographic region, both between and within countries. For example, the density of hospitals in rural regions of South Africa is lower than in urban regions, making accessing them more arduous. The results of the project are,

therefore, likely less applicable to those who access private healthcare and live in rural areas than to those in urban areas accessing public healthcare.

Isolates were collected only from children under five years old in Chapter 3 and the non-South African isolates included in Chapter 5. This means that shigellosis in people five years old and over are un(der)-represented in two studies. However, the South African study results (Chapter 4 and 5) suggest that population clustering by age is minimal, at least in the public healthcare sector, and thus sampling only from those under the age of five (who have the highest prevalence of any group) likely has minimal impact on the scope.

This thesis focuses predominantly on *S. flexneri* 2a and *S. sonnei*, with only one study (Chapter 3) being non-exclusive of serotype, limiting the scope to shigellosis attributable to *S. flexneri* and *S. sonnei*. While other *S. flexneri* serotypes are not included in the bulk of this thesis, the similarity of the epidemiology and pathogen ecology of the different serotypes means that these results can also apply context and understand of the larger *S. flexneri* population. Though more research is needed to confirm that findings hold true for other serotypes.

7.4.2. Future directions

Further research should be done to include or contextualise private healthcare shigellosis in South Africa. It is not currently known if distinct private healthcare-associated strains exist in South Africa or if all strains are present throughout the host population. Without knowledge of private healthcare shigellosis, the picture of shigellosis in South Africa is incomplete. Furthermore, healthcare policy based on the knowledge gained from public healthcare shigellosis may not be relevant to those accessing private healthcare.

This thesis highlights the importance of shigellosis research in sub-Saharan Africa. Further shigellosis research using WGSA on isolates from other sub-Saharan African countries is needed to provide further insights into epidemiology on this region, confirming if the findings from this thesis are broadly

true for shigellosis in the sub-continent, and to create a more complete understanding of the epidemiology of shigellosis globally, particularly as we move toward vaccine licensure which will change *Shigella* population structure in the region.

Future sub-Saharan African shigellosis studies would ideally include many isolates across all *Shigella* serotypes. However, sampling from multiple sites should be a priority to ensure a majority of the population genetic diversity is included, particularly for research on *S. flexneri* due to the highly distinct distribution of strains. The wide diversity and distinct geographical distributions observed in South African *S. flexneri* 2a (Chapter 4), mean that multi-site sampling is required to capture most of the genetic diversity for even a single *S. flexneri* serotype, at a national level. Single site sampling in Malawi (Chapter 3), identified a range of *S. flexneri* serotypes, and some *S. boydii*. However, too few isolates of each serotype were identified to provide any insight into individual serotype epidemiology.

Further research is also needed to better understand the link between HIV and *Shigella* evolution, including the impact of FQR strain emergence. I have hypothesised in this thesis that the observed diversifying evolution is a consequence of increased prevalence of HIV. A case-control study could be used to determine the relative risk of catching shigellosis in HIV+ people compared to HIV- people. Nested within such a study could be a cohort study examining the effects of HIV+ treatment on risk of shigellosis, specifically examining treatment regimen adherence as no patient should be required to abstain from treatment. Such a study would show if being HIV+ increases risk of getting shigellosis, and if treatment decreases risk. Regular testing for shigellosis during the case-control and nested-cohort study would show if HIV has any effect on ability to clear shigellosis infection. With chronic carriage being confirmed via WGSA of the collected samples.

Bibliography

1. Institue for Health Metrics and Evaluation, *Global Burden of Disease.* https://vizhub.healthdata.org/gbd-results/, 2019.
2. Troeger, C.E., et al., *Quantifying risks and interventions that have affected the burden of diarrhoea among children younger than 5 years: an analysis of the Global Burden of Disease Study 2017.* The Lancet Infectious Diseases, 2020. **20**(1): p. 37-59.
3. Paulson, K.R., et al., *Global, regional, and national progress towards Sustainable Development Goal 3.2 for neonatal and child health: all-cause and cause-specific mortality findings from the Global Burden of Disease Study 2019.* The Lancet, 2021. **398**(10303): p. 870-905.
4. Zerbo, A., R. Castro Delgado, and P. Arcos González, *Water sanitation and hygiene in Sub-Saharan Africa: Coverage, risks of diarrheal diseases, and urbanization.* Journal of Biosafety and Biosecurity, 2021. **3**(1): p. 41-45.
5. Coates, M.M., et al., *Burden of disease among the world's poorest billion people: An expert-informed secondary analysis of Global Burden of Disease estimates.* PLOS ONE, 2021. **16**(8): p. e0253073.
6. Belay, D.G., et al., *Open defecation practice and its determinants among households in sub-Saharan Africa: pooled prevalence and multilevel analysis of 33 sub-Saharan Africa countries demographic and health survey.* Tropical Medicine and Health, 2022. **50**(1): p. 28.
7. Reiner, R.C., Jr., et al., *Mapping geographical inequalities in childhood diarrhoeal morbidity and mortality in low-income and middle-income countries, 2000-17: analysis for the Global Burden of Disease Study 2017.* The Lancet, 2020. **395**(10239): p. 1779-1801.
8. *Progress on household drinking water, sanitation and hygiene 2000-2020: Five years into the SDGs.* Geneva: World Health Organization (WHO) and the United Nations Children's Fund (UNICEF), 2021.
9. Wang, H., et al., *Global, regional, and national life expectancy, all-cause mortality, and cause-specific mortality for 249 causes of death, 1980-2015: a systematic analysis for the Global Burden of Disease Study 2015.* The Lancet, 2016. **388**(10053): p. 1459-1544.
10. Khalil, I.A., et al., *Morbidity and mortality due to shigella and enterotoxigenic Escherichia coli diarrhoea: the Global Burden of Disease Study 1990–2016.* Lancet Infectious Disease 2018. **18**: p. 1229 - 1240.
11. Troeger, C., et al., *Estimates of global, regional, and national morbidity, mortality, and aetiologies of diarrhoeal diseases: a systematic analysis for the Global Burden of Disease Study 2015.* Lancet infectious diseases, 2017. **17**: p. 909-948.
12. Hosangadi, D., et al., *WHO consultation on ETEC and Shigella burden of disease, Geneva, 6–7th April 2017: Meeting report.* Vaccine, 2019. **37**(50): p. 7381-7390.
13. Granfors, K., et al., *Bacterial lipopolysaccharide in synovial fluid cells in Shigella triggered reactive arthritis [4].* Journal of Rheumatology, 1992. **19**(3): p. 500.
14. Kotloff, K.L., et al., *Burden and aetiology of diarrhoeal disease in infants and young children in developing countries (the Global Enteric Multicenter Study, GEMS): a prospective, case-control study.* The Lancet, 2013. **382**(9888): p. 209-222.
15. Lee, G., et al., *Effects of shigella-, campylobacter- and ETEC-associated diarrhea on childhood growth.* Pediatric Infectious Disease Journal, 2014. **33**(10): p. 1004-1009.
16. Pinkerton, R., et al., *Early childhood diarrhea predicts cognitive delays in later childhood independently of malnutrition.* American Journal of Tropical Medicine and Hygiene, 2016. **95**(5): p. 1004-1010.
17. Porter, C.K., et al., *Infectious Gastroenteritis and Risk of Developing Inflammatory Bowel Disease.* Gastroenterology, 2008. **135**(3): p. 781-786.

18. Schiellerup, P., K.A. Krogfelt, and H. Locht, *A comparison of self-reported joint symptoms following infection with different enteric pathogens: Effect of HLA-B27.* Journal of Rheumatology, 2008. **35**(3): p. 480-487.
19. Schuster, H.J., et al., *An adult case with shigellosis-associated encephalopathy.* BMJ Case Reports, 2018. **2018**.
20. Kotloff, K.L., et al., *Shigellosis.* The Lancet, 2018. **391**(10122): p. 801-812.
21. WHO, *Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics.* WHO press, 2017: p. 7.
22. WHO, *DRAFT WHO Preferred Product Characteristics for Vaccines against Shigella.* WHO press, 2020.
23. Maurelli, A., et al., *"Black holes" and bacterial pathogenicity: A large genomic deletion that enhances the virulence of Shigella spp. and enteroinvasive Escherichia coli.* Proceedings of the National Academy of Sciences of the United States of America, 1998. **95**: p. 3943-8.
24. McVicker, G. and C.M. Tang, *Deletion of toxin–antitoxin systems in the evolution of Shigella sonnei as a host-adapted pathogen.* Nature Microbiology, 2016. **2**(2): p. 16204.
25. Schroeder, G.N. and H. Hilbi, *Molecular pathogenesis of Shigella spp.: controlling host cell signaling, invasion, and death by type III secretion.* Clin Microbiol Rev, 2008. **21**(1): p. 134-56.
26. Di Martino, M.L., et al., *The Multifaceted Activity of the VirF Regulatory Protein in the Shigella Lifestyle.* Front Mol Biosci, 2016. **3**: p. 61.
27. Corr, S.C., C.C. Gahan, and C. Hill, *M-cells: origin, morphology and role in mucosal immunity and microbial pathogenesis.* FEMS Immunol Med Microbiol, 2008. **52**(1): p. 2-12.
28. Mattock, E. and A.J. Blocker, *How Do the Virulence Factors of Shigella Work Together to Cause Disease?* Front Cell Infect Microbiol, 2017. **7**: p. 64.
29. Seferbekova, Z., et al., *High Rates of Genome Rearrangements and Pathogenicity of Shigella spp.* Front Microbiol, 2021. **12**: p. 628622.
30. Arbab, S., et al., *Drug resistance and susceptibility testing of Gram negative bacterial isolates from healthy cattle with different β - Lactam resistance Phenotypes from Shandong province China.* Braz J Biol, 2021. **83**: p. e247061.
31. Shepherd, J.G., L. Wang, and P.R. Reeves, *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.* Infect Immun, 2000. **68**(10): p. 6056-61.
32. Caboni, M., et al., *An O Antigen Capsule Modulates Bacterial Pathogenesis in Shigella sonnei.* PLOS Pathogens, 2015. **11**(3): p. e1004749.
33. Connor, T.R., et al., *Species-wide whole genome sequencing reveals historical global spread and recent local persistence in Shigella flexneri.* eLife, 2015. **4**(AUGUST2015).
34. Weill, F.X., et al., *Global phylogeography and evolutionary history of Shigella dysenteriae type 1.* Nature Microbiology, 2016. **1**(4).
35. Holt, K.E., et al., *Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe.* Nature genetics, 2012. **44**(9): p. 1056-1059.
36. Kania, D.A., et al., *Genome diversity of Shigella boydii.* Pathog Dis, 2016. **74**(4): p. ftw027.
37. Bengtsson, R.J., et al., *Pathogenomic analyses of Shigella isolates inform factors limiting shigellosis prevention and control across LMICs.* Nature Microbiology, 2022. **7**(2): p. 251-261.
38. Yassine, I., et al., *Population structure analysis and laboratory monitoring of Shigella by core-genome multilocus sequence typing.* Nat Commun, 2022. **13**(1): p. 551.
39. Hawkey, J., et al., *Global population structure and genotyping framework for genomic surveillance of the major dysentery pathogen, Shigella sonnei.* Nature Communications, 2021. **12**(1): p. 2684.
40. Eybpoosh, S., et al., *Molecular epidemiology of infectious diseases.* Electronic physician, 2017. **9**(8): p. 5149-5158.

41. Chang, Y., et al., *Identification of herpesvirus-like DNA sequences in AIDS-associated Kaposi's sarcoma.* Science, 1994. **266**(5192): p. 1865-9.
42. Moore, P.S., et al., *Kaposi's sarcoma-associated herpesvirus infection prior to onset of Kaposi's sarcoma.* Aids, 1996. **10**(2): p. 175-80.
43. Baba, T., et al., *Genome and virulence determinants of high virulence community-acquired MRSA.* Lancet, 2002. **359**(9320): p. 1819-27.
44. Bergeron, C.R., et al., *Chicken as reservoir for extraintestinal pathogenic Escherichia coli in humans, Canada.* Emerging infectious diseases, 2012. **18**(3): p. 415-421.
45. Cortés, P., et al., *Isolation and characterization of potentially pathogenic antimicrobial-resistant Escherichia coli strains from chicken and pig farms in Spain.* Applied and environmental microbiology, 2010. **76**(9): p. 2799-2805.
46. Grassly, N.C. and C. Fraser, *Mathematical models of infectious disease transmission.* Nature Reviews Microbiology, 2008. **6**(6): p. 477-487.
47. Wallinga, J. and M. Lipsitch, *How generation intervals shape the relationship between growth rates and reproductive numbers.* Proc Biol Sci, 2007. **274**(1609): p. 599-604.
48. Islam, A., et al., *Assessment of basic reproduction number (R(0)), spatial and temporal epidemiological determinants, and genetic characterization of SARS-CoV-2 in Bangladesh.* Infect Genet Evol, 2021. **92**: p. 104884.
49. Ferrari, M.J., O.N. Bjørnstad, and A.P. Dobson, *Estimation and inference of R0 of an infectious pathogen by a removal method.* Mathematical Biosciences, 2005. **198**(1): p. 14-26.
50. Sharp, P.M. and B.H. Hahn, *Origins of HIV and the AIDS pandemic.* Cold Spring Harbor perspectives in medicine, 2011. **1**(1): p. a006841-a006841.
51. Slatko, B.E., A.F. Gardner, and F.M. Ausubel, *Overview of Next-Generation Sequencing Technologies.* Curr Protoc Mol Biol, 2018. **122**(1): p. e59.
52. Ewing, B., et al., *Base-calling of automated sequencer traces using phred. I. Accuracy assessment.* Genome research, 1998. **8 3**: p. 175-85.
53. Ewing, B. and P. Green, *Base-calling of automated sequencer traces using phred. II. Error probabilities.* Genome Res, 1998. **8**(3): p. 186-94.
54. Del Fabbro, C., et al., *An extensive evaluation of read trimming effects on Illumina NGS data analysis.* PLoS One, 2013. **8**(12): p. e85024.
55. Bolger, A.M., M. Lohse, and B. Usadel, *Trimmomatic: a flexible trimmer for Illumina sequence data.* Bioinformatics, 2014. **30**(15): p. 2114-20.
56. Pace, N.R., *Microbial phylogeny and evolution: concepts and controversies. Chapter 2. The Large Scale Structure of the Tree of Life.* Oxford University Press, 2005.
57. Ahrenfeldt, J., et al., *Bacterial whole genome-based phylogeny: construction of a new benchmarking dataset and assessment of some existing methods.* BMC Genomics, 2017. **18**(1): p. 19.
58. Schürch, A.C., et al., *Whole genome sequencing options for bacterial strain typing and epidemiologic analysis based on single nucleotide polymorphism versus gene-by-gene–based approaches.* Clinical Microbiology and Infection, 2018. **24**(4): p. 350-354.
59. Saltykova, A., et al., *Detailed Evaluation of Data Analysis Tools for Subtyping of Bacterial Isolates Based on Whole Genome Sequencing: Neisseria meningitidis as a Proof of Concept.* Front Microbiol, 2019. **10**: p. 2897.
60. Xia, X., *Maximum Likelihood in Molecular Phylogenetics*, in *Bioinformatics and the Cell: Modern Computational Approaches in Genomics, Proteomics and Transcriptomics*, X. Xia, Editor. 2018, Springer International Publishing: Cham. p. 381-395.
61. Stamatakis, A., *RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies.* Bioinformatics, 2014. **30**(9): p. 1312-1313.
62. Kozlov, A.M., et al., *RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference.* Bioinformatics, 2019. **35**(21): p. 4453-4455.

63. Felsenstein, J., *Confidence limits on phylogenies: an approach using the bootstrap.* Evolution, 1985. **39**(4): p. 783-791.
64. Sullivan, J., *Maximum-Likelihood Methods for Phylogeny Estimation*, in *Methods in Enzymology*. 2005, Academic Press. p. 757-779.
65. Oaks, J.R., *Bayesian Phylogenetics: Methods, Algorithms, and Applications. — Edited by Ming-Hui Chen, Lynn Kuo, and Paul O. Lewis*. 2015, Oxford University Press: Great Britain. p. 1122-1125.
66. Downey, A., *Think Bayes*. 2013: O'Reilly.
67. Lan, R. and P.R. Reeves, *Escherichia coli in disguise: molecular origins of Shigella.* Microbes and Infection, 2002. **4**(11): p. 1125-1132.
68. Lan, R., et al., *Molecular Evolutionary Relationships of Enteroinvasive Escherichia coli and Shigella spp.* Infection and Immunity, 2004. **72**(9): p. 5080.
69. Sun, Q., et al., *Identification and Characterization of a Novel Shigella flexneri Serotype Yv in China.* PLOS ONE, 2013. **8**(7): p. e70238.
70. Huan, P.T., et al., *Molecular characterization of the genes involved in O-antigen modification, attachment, integration and excision in Shigella flexneri bacteriophage SfV.* Gene, 1997. **195**(2): p. 217-27.
71. Knirel, Y.A., et al., *O-antigen modifications providing antigenic diversity of Shigella flexneri and underlying genetic mechanisms.* Biochemistry (Mosc), 2015. **80**(7): p. 901-14.
72. Baker, K.S., et al., *Horizontal antimicrobial resistance transfer drives epidemics of multiple Shigella species.* Nat Commun, 2018. **9**(1): p. 1462.
73. Holt, K.E., et al., *Tracking the establishment of local endemic populations of an emergent enteric pathogen.* Proc Natl Acad Sci U S A, 2013. **110**(43): p. 17522-7.
74. Kotloff, K.L., et al., *Global burden of Shigella infections: implications for vaccine development and implementation of control strategies.* Bull World Health Organ, 1999. **77**(8): p. 651-66.
75. Sayeed, S., D.A. Sack, and F. Qadri, *Protection from Shigella sonnei infection by immunisation of rabbits with Plesiomonas shigelloides (SVC 01).* J Med Microbiol, 1992. **37**(6): p. 382-4.
76. Chung The, H., et al., *The genomic signatures of Shigella evolution, adaptation and geographical spread.* Nature Reviews Microbiology, 2016. **14**(4): p. 235-250.
77. Pilla, G., G. McVicker, and C.M. Tang, *Genetic plasticity of the Shigella virulence plasmid is mediated by intra- and inter-molecular events between insertion sequences.* PLoS Genet, 2017. **13**(9): p. e1007014.
78. Chung The, H. and S. Baker, *Out of Asia: the independent rise and global spread of fluoroquinolone-resistant Shigella.* Microb Genom, 2018. **4**(4).
79. Chung The, H., et al., *South Asia as a Reservoir for the Global Spread of Ciprofloxacin-Resistant Shigella sonnei: A Cross-Sectional Study.* PLoS Medicine, 2016. **13**(8).
80. Nüesch-Inderbinen, M., et al., *Shigella Antimicrobial Drug Resistance Mechanisms, 2004-2014.* Emerging infectious diseases, 2016. **22**(6): p. 1083-1085.
81. Chung The, H., et al., *Evolutionary histories and antimicrobial resistance in Shigella flexneri and Shigella sonnei in Southeast Asia.* Communications Biology, 2021. **4**(1): p. 353.
82. Rajakumar, K., et al., *Identification of a Chromosomal Shigella flexneri Multi-Antibiotic Resistance Locus Which Shares Sequence and Organizational Similarity with the Resistance Region of the Plasmid NR1.* Plasmid, 1997. **37**(3): p. 159-168.
83. Luck, S.N., et al., *Ferric dicitrate transport system (Fec) of Shigella flexneri 2a YSH6000 is encoded on a novel pathogenicity island carrying multiple antibiotic resistance genes.* Infect Immun, 2001. **69**(10): p. 6012-21.
84. van den Beld, M.J. and F.A. Reubsaet, *Differentiation between Shigella, enteroinvasive Escherichia coli (EIEC) and noninvasive Escherichia coli.* Eur J Clin Microbiol Infect Dis, 2012. **31**(6): p. 899-904.
85. Bengtsson, R.J., et al., *Accessory Genome Dynamics and Structural Variation of Shigella from Persistent Infections.* mBio, 2021. **12**(2).

86. Hussen, S., G. Mulatu, and Z. Yohannes Kassa, *Prevalence of Shigella species and its drug resistance pattern in Ethiopia: a systematic review and meta-analysis.* Annals of Clinical Microbiology and Antimicrobials, 2019. **18**(1): p. 22.
87. Mero, S., et al., *Prevalence of diarrhoeal pathogens among children under five years of age with and without diarrhoea in Guinea-Bissau.* PLoS Negl Trop Dis, 2021. **15**(9): p. e0009709.
88. Becker, S.L., et al., *Combined stool-based multiplex PCR and microscopy for enhanced pathogen detection in patients with persistent diarrhoea and asymptomatic controls from Cote d'Ivoire.* Clinical Microbiology and Infection, 2015. **21**(6): p. 591.e1-591.e10.
89. Veronica Di, C., et al., *Application of Luminex Gastrointestinal Pathogen Panel to human stool samples from Côte d'Ivoire.* The Journal of Infection in Developing Countries, 2015. **9**(08).
90. Akuffo, R., et al., *Prevalence of enteric infections among hospitalized patients in two referral hospitals in Ghana.* BMC Research Notes, 2017. **10**(1).
91. Dzotsi, E.K., et al., *Surveillance of Bacterial Pathogens of Diarrhoea in Two Selected Sub Metros Within the Accra Metropolis.* Ghana medical journal, 2015. **49**(2): p. 65-71.
92. Krumkamp, R., et al., *Gastrointestinal Infections and Diarrheal Disease in Ghanaian Infants and Children: An Outpatient Case-Control Study.* PLOS Neglected Tropical Diseases, 2015. **9**(3): p. e0003568.
93. Simpore, J., et al., *Aetiology of acute gastro-enteritis in children at saint camille medical centre, ouagadougou, Burkina Faso.* Pakistan Journal of Biological Sciences, 2009. **12**(3): p. 258-263.
94. Bradbury, R.S., et al., *Enteric pathogens of food sellers in rural Gambia with incidental finding of Myxobolus species (Protozoa: Myxozoa).* Transactions of The Royal Society of Tropical Medicine and Hygiene, 2015. **109**(5): p. 334-339.
95. Ali Nor, B.S., N.C. Menza, and A.M. Musyoki, *Multidrug-Resistant Shigellosis among Children Aged below Five Years with Diarrhea at Banadir Hospital in Mogadishu, Somalia.* Canadian Journal of Infectious Diseases and Medical Microbiology, 2021. **2021**: p. 6630272.
96. Moremi, N., et al., *Prevalence and antimicrobial sensitivity of Shiga-toxin-producing Escherichia coli among underfives presenting with diarrhoea at hospitals in Mwanza City Tanzania.* Tanzania journal of health research, 2017. **19**.
97. Platts-Mills, J.A., et al., *Pathogen-specific burdens of community diarrhoea in developing countries: a multisite birth cohort study (MAL-ED).* The Lancet Global Health, 2015. **3**(9): p. e564-e575.
98. Platts-Mills, J.A., et al., *Impact of Rotavirus Vaccine Introduction and Postintroduction Etiology of Diarrhea Requiring Hospital Admission in Haydom, Tanzania, a Rural African Setting.* Clinical Infectious Diseases, 2017. **65**(7): p. 1144-1151.
99. Zachariah, O.H., et al., *Multiple drug resistance of Campylobacter jejuni and Shigella isolated from diarrhoeic children at Kapsabet County referral hospital, Kenya.* BMC Infectious Diseases, 2021. **21**(1): p. 109.
100. Oliver, W.M., et al., *Etiology and pathogenicity of bacterial isolates: a cross sectional study among diarrheal children below five years in central regions of Kenya.* PAMJ, 2018. **31**(88).
101. Kilongosi Webale, M., et al., *Epidemiological patterns and antimicrobial resistance of bacterial diarrhea among children in Nairobi City, Kenya.* Gastroenterology and Hepatology: from Bed to Bench, 2020. **12**(3): p. 8.
102. Pavlinac, P.B., et al., *Failure of Syndrome-Based Diarrhea Management Guidelines to Detect Shigella Infections in Kenyan Children.* Journal of the Pediatric Infectious Diseases Society, 2015. **5**(4): p. 366-374.
103. Shah, M., et al., *Prevalence, seasonal variation, and antibiotic resistance pattern of enteric bacterial pathogens among hospitalized diarrheic children in suburban regions of central Kenya.* Tropical Medicine and Health, 2016. **44**(1): p. 39.
104. Nyanga, P.L., et al., *Escherichia coli pathotypes and Shigella sero-groups in diarrheic children in Nairobi city, Kenya.* Gastroenterol Hepatol Bed Bench, 2017. **10**(3): p. 220-228.

105. Iturriza-Gomara, M., et al., *Etiology of Diarrhea Among Hospitalized Children in Blantyre, Malawi, Following Rotavirus Vaccine Introduction: A Case-Control Study.* J Infect Dis, 2019. **220**(2): p. 213-218.
106. Phiri, A.F.N.D., et al., *Burden, Antibiotic Resistance, and Clonality of Shigella spp. Implicated in Community-Acquired Acute Diarrhoea in Lilongwe, Malawi.* Tropical Medicine and Infectious Disease, 2021. **6**(2): p. 63.
107. Randremanana, R.V., et al., *Etiologies, Risk Factors and Impact of Severe Diarrhea in the Under-Fives in Moramanga and Antananarivo, Madagascar.* PLOS ONE, 2016. **11**(7): p. e0158862.
108. Nhampossa, T., et al., *Diarrheal Disease in Rural Mozambique: Burden, Risk Factors and Etiology of Diarrheal Disease among Children Aged 0–59 Months Seeking Care at Health Facilities.* PLOS ONE, 2015. **10**(5): p. e0119824.
109. Vubil, D., et al., *Clinical features, risk factors, and impact of antibiotic treatment of diarrhea caused by Shigella in children less than 5 years in Manhiça District, rural Mozambique.* Infect Drug Resist, 2018. **11**: p. 2095-2106.
110. Adam, M.A., et al., *Molecular Survey of Viral and Bacterial Causes of Childhood Diarrhea in Khartoum State, Sudan.* Frontiers in Microbiology, 2018. **9**.
111. Saeed, A., H. Abd, and G. Sandstrom, *Microbial aetiology of acute diarrhoea in children under five years of age in Khartoum, Sudan.* Journal of Medical Microbiology, 2015. **64**(4): p. 432-437.
112. Chiyangi, H., et al., *Identification and antimicrobial resistance patterns of bacterial enteropathogens from children aged 0–59 months at the University Teaching Hospital, Lusaka, Zambia: a prospective cross sectional study.* BMC Infectious Diseases, 2017. **17**(1): p. 117.
113. Bliss, J., et al., *High Prevalence of Shigella or Enteroinvasive Escherichia coli Carriage among Residents of an Internally Displaced Persons Camp in South Sudan.* The American Journal of Tropical Medicine and Hygiene, 2018. **98**(2): p. 595-597.
114. Irenge, L.M., et al. *Antimicrobial resistance of bacteria isolated from patients with bloodstream infections at a tertiary care hospital in the Democratic Republic of the Congo*. South African medical journal = Suid-Afrikaanse tydskrif vir geneeskunde, 2015. **105**, 752-755 DOI: 10.7196/samjnew.7937.
115. Pernica, J.M., et al., *Rapid enteric testing to permit targeted antimicrobial therapy, with and without Lactobacillus reuteri probiotics, for paediatric acute diarrhoeal disease in Botswana: A pilot, randomized, factorial, controlled trial.* PLOS ONE, 2017. **12**(10): p. e0185177.
116. Mokomane, M., et al., *A comparison of flocked swabs and traditional swabs, using multiplex real-time PCR for detection of common gastroenteritis pathogens in Botswana.* Diagnostic Microbiology and Infectious Disease, 2016. **86**(2): p. 141-143.
117. Pernica, J.M., et al., *Correlation of Clinical Outcomes With Multiplex Molecular Testing of Stool From Children Admitted to Hospital With Gastroenteritis in Botswana.* Journal of the Pediatric Infectious Diseases Society, 2015. **5**(3): p. 312-318.
118. Langendorf, C., et al., *Enteric Bacterial Pathogens in Children with Diarrhea in Niger: Diversity and Antimicrobial Resistance.* PLOS ONE, 2015. **10**(3): p. e0120275.
119. Wang, H., et al., *A Prospective Study of Etiological Agents Among Febrile Patients in Sierra Leone.* Infectious Diseases and Therapy, 2021. **10**(3): p. 1645-1664.
120. Tosisa, W., et al., *Prevalence and antimicrobial susceptibility of Salmonella and Shigella species isolated from diarrheic children in Ambo town.* BMC Pediatrics, 2020. **20**(1): p. 91.
121. Njuguna, C., et al., *Enteric pathogens and factors associated with acute bloody diarrhoea, Kenya.* BMC Infectious Diseases, 2016. **16**(1): p. 477.
122. Chissaque, A., et al., *The Epidemiology of Diarrhea in Children Under 5 Years of Age in Mozambique.* Current Tropical Medicine Reports, 2018. **5**(3): p. 115-124.

123. Vubil, D., et al., *Antibiotic resistance and molecular characterization of shigella isolates recovered from children aged less than 5 years in Manhiça, Southern Mozambique.* International Journal of Antimicrobial Agents, 2018. **51**(6): p. 881-887.
124. Breurec, S., et al., *Serotype Distribution and Antimicrobial Resistance of Shigella Species in Bangui, Central African Republic, from 2002 to 2013.* The American Journal of Tropical Medicine and Hygiene, 2018. **99**(2): p. 283-286.
125. Schaumburg, F., et al., *Molecular characterization of Shigella spp. from patients in Gabon 2011–2013.* Transactions of The Royal Society of Tropical Medicine and Hygiene, 2015. **109**(4): p. 275-279.
126. Kalule, J.B., et al., *Prevalence and antibiotic susceptibility patterns of enteric bacterial pathogens in human and non-human sources in an urban informal settlement in Cape Town, South Africa.* BMC Microbiology, 2019. **19**(1): p. 244.
127. Kahsay, A.G. and S. Muthupandian, *A review on Sero diversity and antimicrobial resistance patterns of Shigella species in Africa, Asia and South America, 2001–2014.* BMC Research Notes, 2016. **9**(1): p. 422.
128. Bar-Zeev, N., et al., *Effectiveness of a monovalent rotavirus vaccine in infants in Malawi after programmatic roll-out: an observational and case-control study.* The Lancet Infectious Diseases, 2015. **15**(4): p. 422-428.
129. Kotloff, K.L., et al., *The Global Enteric Multicenter Study (GEMS) of diarrheal disease in infants and young children in developing countries: epidemiologic and clinical methods of the case/control study.* Clin Infect Dis, 2012. **55 Suppl 4**: p. S232-45.
130. Latif, H., et al., *A Gapless, Unambiguous Genome Sequence of the Enterohemorrhagic Escherichia coli O157:H7 Strain EDL933.* Genome Announc, 2014. **2**(4).
131. Mehta, H.H., et al., *Genome-wide analysis of the response to nitric oxide in uropathogenic Escherichia coli CFT073.* Microb Genom, 2015. **1**(4): p. e000031.
132. Ashton, P.M., et al., *Draft genome sequences of the type strains of Shigella flexneri held at Public Health England: comparison of classical phenotypic and novel molecular assays with whole genome sequence.* Gut Pathogens, 2014. **6**(1): p. 7.
133. Baker, K.S., et al., *Intercontinental dissemination of azithromycin-resistant shigellosis through sexual transmission: A cross-sectional study.* The Lancet Infectious Diseases, 2015. **15**(8): p. 913-921.
134. Baker, K.S., et al., *The extant World War 1 dysentery bacillus NCTC1: a genomic analysis.* Lancet, 2014. **384**(9955): p. 1691-7.
135. Quail, M.A., et al., *Optimal enzymes for amplifying sequencing libraries.* Nature Methods, 2012. **9**(1): p. 10-11.
136. Perez-Sepulveda, B.M., et al., *An accessible, efficient and global approach for the large-scale sequencing of bacterial genomes.* Genome Biology, 2021. **22**(1): p. 349.
137. Andrews, S., *FastQC A Quality Control tool for High Throughput Sequence Data*. 2014.
138. Ewels, P., et al., *MultiQC: summarize analysis results for multiple tools and samples in a single report.* Bioinformatics, 2016. **32**(19): p. 3047-8.
139. Okonechnikov, K., A. Conesa, and F. Garcia-Alcalde, *Qualimap 2: advanced multi-sample quality control for high-throughput sequencing data.* Bioinformatics, 2016. **32**(2): p. 292-4.
140. Gurevich, A., et al., *QUAST: quality assessment tool for genome assemblies.* Bioinformatics, 2013. **29**(8): p. 1072-1075.
141. Wick, R.R., et al., *Bandage: interactive visualization of de novo genome assemblies.* Bioinformatics, 2015. **31**(20): p. 3350-3352.
142. Li, H., et al., *The Sequence Alignment/Map format and SAMtools.* Bioinformatics, 2009. **25**(16): p. 2078-9.
143. Li, H., *A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data.* Bioinformatics, 2011. **27**(21): p. 2987-93.

144. Danecek, P., et al., *Twelve years of SAMtools and BCFtools.* GigaScience, 2021. **10**(2).
145. Quinlan, A.R. and I.M. Hall, *BEDTools: a flexible suite of utilities for comparing genomic features.* Bioinformatics (Oxford, England), 2010. **26**(6): p. 841-842.
146. Arndt, D., et al., *PHASTER: a better, faster version of the PHAST phage search tool.* Nucleic Acids Res, 2016. **44**(W1): p. W16-21.
147. Zhou, Y., et al., *PHAST: a fast phage search tool.* Nucleic Acids Res, 2011. **39**(Web Server issue): p. W347-52.
148. Croucher, N.J., et al., *Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins.* Nucleic acids research, 2015. **43**(3): p. e15.
149. Wick, R.R., et al., *Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads.* PLoS Comput Biol, 2017. **13**(6): p. e1005595.
150. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2011.* Available at: http://www.nicd.ac.za/assets/files/2011%20GERMS-SA%20Annual%20report%20pub%20final(1).pdf, 2011.
151. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2012.* Available at: http://www.nicd.ac.za/assets/files/GERMS-SA%202012%20Annual%20Report.pdf, 2012.
152. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2013.* Available at: http://www.nicd.ac.za/assets/files/GERMS-SA%20AR%202013(1).pdf, 2013.
153. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2014.* Available at: http://www.nicd.ac.za/assets/files/GERMS-SA%20AR%202014(1).pdf, 2014.
154. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2015.* Available at: http://www.nicd.ac.za/assets/files/GERMS-SA%20AR%202015-1.pdf, 2015.
155. Feldgarden, M., et al., *AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence.* Scientific reports, 2021. **11**(1): p. 12728-12728.
156. Zankari, E., et al., *Identification of acquired antimicrobial resistance genes.* The Journal of antimicrobial chemotherapy, 2012. **67**(11): p. 2640-2644.
157. Seemann, T., *Prokka: rapid prokaryotic genome annotation.* (1367-4811 (Electronic)).
158. Letunic, I. and P. Bork, *Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation.* Nucleic Acids Research, 2021. **49**(W1): p. W293-W296.
159. Stenhouse, G.E., et al., *Whole genome sequence analysis of Shigella from Malawi identifies fluoroquinolone resistance.* Microb Genom, 2021. **7**(5).
160. Siguier, P., et al., *ISfinder: the reference centre for bacterial insertion sequences.* Nucleic acids research, 2006. **34**(Database issue): p. D32-6.
161. Vinh, H., et al., *A changing picture of shigellosis in southern Vietnam: shifting species dominance, antimicrobial susceptibility and clinical presentation.* BMC Infectious Diseases, 2009. **9**(1): p. 204.
162. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2007.* Available at: https://www.nicd.ac.za/assets/files/2007_GERMS-SA_Annual_Report.pdf, 2007.
163. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2008.* Available at: https://www.nicd.ac.za/assets/files/2008_GERMS-SA_Annual_Report(1).pdf, 2008.
164. Group for Enteric, R.a.M.d.S.i.S.A., *GERMS-SA Annual Report 2009.* Available at: https://www.nicd.ac.za/assets/files/2009GERMS-SA_Annual_Report.pdf, 2009.
165. Rambaut, A., et al., *Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen).* Virus Evol, 2016. **2**(1): p. vew007.
166. Rambaut, A., et al., *Bayesian Coalescent Inference of Past Population Dynamics from Molecular Sequences.* Molecular Biology and Evolution, 2005. **22**(5): p. 1185-1192.
167. Bouckaert, R.R. and A.J. Drummond, *bModelTest: Bayesian phylogenetic site model averaging and model comparison.* BMC Evolutionary Biology, 2017. **17**(1): p. 42.
168. Rambaut, A., et al., *Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7.* Systematic Biology, 2018. **67**(5): p. 901-904.

169. Cheng, L., et al., *Hierarchical and Spatially Explicit Clustering of DNA Sequences with BAPS Software.* Molecular Biology and Evolution, 2013. **30**: p. 1224 - 1228.
170. Tonkin-Hill, G., et al., *RhierBAPS: An R implementation of the population clustering algorithm hierBAPS [version 1; peer review: 2 approved].* Wellcome Open Research, 2018. **3**(93).
171. Shen, W., et al., *SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation.* PLOS ONE, 2016. **11**(10): p. e0163962.
172. Heled, J. and A.J. Drummond, *Bayesian inference of population size history from multiple loci.* BMC Evolutionary Biology, 2008. **8**(1): p. 289.
173. Müller, N.F., D. Rasmussen, and T. Stadler, *MASCOT: parameter and state inference under the marginal structured coalescent approximation.* Bioinformatics, 2018. **34**(22): p. 3843-3848.
174. Müller, N.F., D.A. Rasmussen, and T. Stadler, *The Structured Coalescent and Its Approximations.* Mol Biol Evol, 2017. **34**(11): p. 2970-2981.
175. Vaughan, T.G., et al., *Efficient Bayesian inference under the structured coalescent.* Bioinformatics, 2014. **30**(16): p. 2272-2279.
176. Kim, H.Y., *Statistical notes for clinical researchers: Chi-squared test and Fisher's exact test.* Restor Dent Endod, 2017. **42**(2): p. 152-155.
177. Stanberry, L., *Exact Test for Independence*, in *Encyclopedia of Systems Biology*, W. Dubitzky, et al., Editors. 2013, Springer New York: New York, NY. p. 697-698.
178. Haynes, W., *Bonferroni Correction*, in *Encyclopedia of Systems Biology*, W. Dubitzky, et al., Editors. 2013, Springer New York: New York, NY. p. 154-154.
179. Boll, E.J., et al., *The Fimbriae of Enteroaggregative Escherichia coli Induce Epithelial Inflammation In Vitro and in a Human Intestinal Xenograft Model.* The Journal of Infectious Diseases, 2012. **206**(5): p. 714-722.
180. Chanin, R.B., et al., *Shigella flexneri Adherence Factor Expression in In Vivo-Like Conditions.* mSphere, 2019. **4**(6).
181. Schwan, W.R., *Regulation of fim genes in uropathogenic Escherichia coli.* World J Clin Infect Dis, 2011. **1**(1): p. 17-25.
182. Schmitt, M.P. and S.M. Payne, *Genetic analysis of the enterobactin gene cluster in Shigella flexneri.* Journal of Bacteriology, 1991. **173**(2): p. 816-825.
183. Schmitt, M.P. and S.M. Payne, *Genetics and regulation of enterobactin genes in Shigella flexneri.* Journal of Bacteriology, 1988. **170**(12): p. 5579-5587.
184. Consortium, T.U., *UniProt: the universal protein knowledgebase in 2021.* Nucleic Acids Research, 2020. **49**(D1): p. D480-D489.
185. Casalino, M., et al., *Characterization of endemic Shigella flexneri strains in Somalia: antimicrobial resistance, plasmid profiles, and serotype correlation.* J Clin Microbiol, 1994. **32**(5): p. 1179-83.
186. Pulford, C.V., et al., *Stepwise evolution of Salmonella Typhimurium ST313 causing bloodstream infection in Africa.* Nature microbiology, 2021. **6**(3): p. 327-338.
187. UNAIDS, *AIDS info: epidemic and response.* https://aidsinfo.unaids.org/, 2021.
188. Aragón, T.J., et al., *Case-control study of shigellosis in San Francisco: The role of sexual transmission and HIV infection.* Clinical Infectious Diseases, 2007. **44**(3): p. 327-334.
189. Gilbart, V.L., et al., *Sex, drugs and smart phone applications: Findings from semistructured interviews with men who have sex with men diagnosed with Shigella flexneri 3a in England and Wales.* Sexually Transmitted Infections, 2015. **91**(8): p. 598-602.
190. Küstner, H.G., J.P. Swanevelder, and A. Van Middelkoop, *National HIV surveillance--South Africa, 1990-1992.* S Afr Med J, 1994. **84**(4): p. 195-200.
191. Williams, B. and C. Campbell, *Understanding the epidemic of HIV in South Africa. Analysis of the antenatal clinic survey data.* S Afr Med J, 1998. **88**(3): p. 247-51.
192. Womble, D.D. and R.H. Rownd, *Genetic and physical map of plasmid NR1: comparison with other IncFII antibiotic resistance plasmids.* Microbiological Reviews, 1988. **52**(4): p. 433-451.

193. Turner, S.A., et al., *Nested deletions of the SRL pathogenicity island of Shigella flexneri 2a.* J Bacteriol, 2001. **183**(19): p. 5535-43.
194. Toro, C.S., et al., *Antimicrobial Resistance Dynamics in Chilean Shigella sonnei Strains Within Two Decades: Role of Shigella Resistance Locus Pathogenicity Island and Class 1 and Class 2 Integrons.* Frontiers in Microbiology, 2022. **12**.
195. Agüero, M.E., et al., *A plasmid-encoded outer membrane protein, TraT, enhances resistance of Escherichia coli to phagocytosis.* Infect Immun, 1984. **46**(3): p. 740-6.
196. Moll, A., P.A. Manning, and K.N. Timmis, *Plasmid-determined resistance to serum bactericidal activity: a major outer membrane protein, the traT gene product, is responsible for plasmid-specified serum resistance in Escherichia coli.* Infect Immun, 1980. **28**(2): p. 359-67.
197. Kanukollu, U., et al., *Contribution of the traT gene to serum resistance among clinical isolates of enterobacteriaceae.* J Med Microbiol, 1985. **19**(1): p. 61-7.
198. Fujiyama, R., et al., *The shf Gene of a Shigella flexneri Homologue on the Virulent Plasmid pAA2 of Enteroaggregative Escherichia coli 042 Is Required for Firm Biofilm Formation.* Current Microbiology, 2008. **56**(5): p. 474-480.
199. Czeczulin, J.R., et al., *Phylogenetic analysis of enteroaggregative and diffusely adherent Escherichia coli.* Infect Immun, 1999. **67**(6): p. 2692-9.
200. Keddy, K.H., et al., *Systemic shigellosis in South Africa.* Clinical Infectious Diseases, 2012. **54**(10): p. 1448-1454.
201. Sansonetti, P.J., D.J. Kopecko, and S.B. Formal, *Shigella sonnei plasmids: evidence that a large plasmid is necessary for virulence.* Infection and immunity, 1981. **34**(1): p. 75-83.
202. Wu, Y., et al., *In Silico Serotyping Based on Whole-Genome Sequencing Improves the Accuracy of Shigella Identification.* Applied and Environmental Microbiology, 2019. **85**(7): p. e00165-19.
203. Li, H., *Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.* arXiv:1303.3997v1 [q-bio.GN], 2013.
204. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows-Wheeler transform.* Bioinformatics, 2009. **25**(14): p. 1754-60.
205. Beraud, M., et al., *A proteomic analysis reveals differential regulation of the σ(S)-dependent yciGFE(katN) locus by YncC and H-NS in Salmonella and Escherichia coli K-12.* Mol Cell Proteomics, 2010. **9**(12): p. 2601-16.
206. Baker, K.S., et al., *Genomic epidemiology of Shigella in the United Kingdom shows transmission of pathogen sublineages and determinants of antimicrobial resistance.* Scientific Reports, 2018. **8**(1).
207. Gu, B., et al., *A 10-year surveillance of antimicrobial susceptibility patterns in Shigella sonnei isolates circulating in Jiangsu Province, China.* Journal of Global Antimicrobial Resistance, 2017. **10**: p. 29-34.
208. Ud-Din, A.I.M.S., et al., *Changing Trends in the Prevalence of Shigella Species: Emergence of Multi-Drug Resistant Shigella sonnei Biotype g in Bangladesh.* PLOS ONE, 2013. **8**(12): p. e82601.
209. Vrints, M., et al., *Surveillance of antibiotic susceptibility patterns among Shigella sonnei strains isolated in Belgium during the 18-year period 1990 to 2007.* J Clin Microbiol, 2009. **47**(5): p. 1379-85.
210. Wang, Y., et al., *Antimicrobial resistance and genetic characterization of Shigella spp. in Shanxi Province, China, during 2006–2016.* BMC Microbiology, 2019. **19**(1): p. 116.
211. Baker, K.S., et al., *Whole genome sequencing of Shigella sonnei through PulseNet Latin America and Caribbean: advancing global surveillance of foodborne illnesses.* Clin Microbiol Infect, 2017. **23**(11): p. 845-853.
212. Abelman, R.L., et al., *Use of whole genome sequencing in surveillance for antimicrobial-resistant Shigella sonnei infections acquired from domestic and international sources.* Microbial Genomics, 2019. **5**(5).

213. Zhao, Z., et al., *Relative transmissibility of shigellosis among different age groups: A modeling study in Hubei Province, China.* PLOS Neglected Tropical Diseases, 2021. **15**(6): p. e0009501.
214. Zhao, Z.Y., et al., *Relative transmissibility of shigellosis among male and female individuals: a modeling study in Hubei Province, China.* Infect Dis Poverty, 2020. **9**(1): p. 39.
215. Roehrich, A.D., et al., *Shigella IpaD has a dual role: signal transduction from the type III secretion system needle tip and intracellular secretion regulation.* Molecular microbiology, 2013. **87**(3): p. 690-706.
216. Mader, A., et al., *Amount of Colicin Release in Escherichia coli Is Regulated by Lysis Gene Expression of the Colicin E2 Operon.* PLOS ONE, 2015. **10**(3): p. e0119124.
217. Cascales, E., et al., *Colicin biology.* Microbiol Mol Biol Rev, 2007. **71**(1): p. 158-229.
218. Narasimhan, V., et al., *BCFtools/RoH: a hidden Markov model approach for detecting autozygosity from next-generation sequencing data.* Bioinformatics (Oxford, England), 2016. **32**(11): p. 1749-1751.
219. Corander, J., et al., *Corander J, Marttinen P, Siren J, Tang J.. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. BMC Bioinform 9: 539*. Vol. 9. 2009. 539.
220. Davies, N.E.C.G. and A.S. Karstaedt, *Shigella bacteraemia over a decade in Soweto, South Africa.* Transactions of the Royal Society of Tropical Medicine and Hygiene, 2008. **102**(12): p. 1269-1273.
221. Serafino Wani, R.L., et al., *Invasive shigellosis in MSM.* International Journal of STD and AIDS, 2016. **27**(10): p. 917-919.
222. Sethuvel, D.P.M., et al., *Phylogenetic and Evolutionary Analysis Reveals the Recent Dominance of Ciprofloxacin-Resistant Shigella sonnei and Local Persistence of S. flexneri Clones in India.* mSphere, 2020. **5**(5): p. e00569-20.
223. Hoffmann, C., et al., *High rates of quinolone-resistant strains of Shigella sonnei in HIV-infected MSM.* Infection, 2013. **41**(5): p. 999-1003.
224. Eikmeier, D., et al., *Decreased Susceptibility to Azithromycin in Clinical Shigella Isolates Associated with HIV and Sexually Transmitted Bacterial Diseases, Minnesota, USA, 2012-2015.* Emerg Infect Dis, 2020. **26**(4): p. 667-674.
225. Richardson, D., et al., *Sexually transmitted Shigella flexneri and ;Shigella sonnei in men who have sex with men.* Sexually Transmitted Infections, 2021. **97**(3): p. 244.
226. Tsai, C.-S., et al., *Changing epidemiology of shigellosis in Taiwan, 2010-2019: an emerging threat to HIV-infected patients and men who have sex with men.* Emerging Microbes & Infections, 2022. **11**(1): p. 498-506.
227. Bardsley, M., et al., *Persistent Transmission of Shigellosis in England Is Associated with a Recently Emerged Multidrug-Resistant Strain of Shigella sonnei.* Journal of Clinical Microbiology, 2020. **58**(4): p. e01692-19.
228. Simms, I., et al., *Intensified shigellosis epidemic associated with sexual transmission in men who have sex with men - Shigella fexneri and s. Sonnei in England, 2004 to end of February 2015.* Eurosurveillance: bulletin europeen sur les maladies transmissibles = European communicable disease bulletin, 2015. **20**.
229. Borg, M.L., et al., *Ongoing outbreak of Shigella flexneri serotype 3a in men who have sex with men in England and Wales, data from 2009–2011.* Eurosurveillance, 2012. **17**(13).
230. Bovée, L.P.M.J., P.G.H. Peerbooms, and J.A.R. Van Den Hoek, *Shigellosis, a sexually transmitted disease in homosexual men.* Nederlands Tijdschrift voor Geneeskunde, 2003. **147**(49): p. 2438-2439.
231. Mohan, K., et al., *What is the overlap between HIV and shigellosis epidemics in England: further evidence of MSM transmission?* Sex Transm Infect, 2018. **94**(1): p. 67-71.
232. Wu, H.H., et al., *Shigellosis outbreak among MSM living with HIV: A case-control study in Taiwan, 2015-2016.* Sexually Transmitted Infections, 2018.

233. Daskalakis, D.C. and M.J. Blaser, *Another perfect storm: Shigella, men who have sex with men, and HIV.* Clinical Infectious Diseases, 2007. **44**(3): p. 335-337.
234. Chung The, H., et al., *Dissecting the molecular evolution of fluoroquinolone-resistant Shigella sonnei.* Nat Commun, 2019. **10**(1): p. 4828.
235. World Bank Group, *World Bank Open Data.* https://data.worldbank.org/, 2020.
236. Chung The, H., et al., *South Asia as a Reservoir for the Global Spread of Ciprofloxacin-Resistant Shigella sonnei: A Cross-Sectional Study.* PLOS Medicine, 2016. **13**(8): p. e1002055.
237. Pilla, G. and C.M. Tang, *Going around in circles: virulence plasmids in enteric pathogens.* Nature Reviews Microbiology, 2018. **16**(8): p. 484-495.
238. Lobato-Márquez, D., et al., *Stabilization of the Virulence Plasmid pSLT of Salmonella Typhimurium by Three Maintenance Systems and Its Evaluation by Using a New Stability Test.* Front Mol Biosci, 2016. **3**: p. 66.
239. Jechalke, S., et al., *Salmonella Establishment in Agricultural Soil and Colonization of Crop Plants Depend on Soil Type and Plant Species.* Frontiers in Microbiology, 2019. **10**.
240. Branchu, P., et al., *Genome Variation and Molecular Epidemiology of Salmonella enterica Serovar Typhimurium Pathovariants.* Infection and Immunity, 2018. **86**(8): p. e00079-18.
241. Hiley, L., R.M.A. Graham, and A.V. Jennison, *Genetic characterisation of variants of the virulence plasmid, pSLT, in Salmonella enterica serovar Typhimurium provides evidence of a variety of evolutionary directions consistent with vertical rather than horizontal transmission.* PLOS ONE, 2019. **14**(4): p. e0215207.
242. Bhaduri, S., *Effect of fat in ground beef on the growth and virulence plasmid (pYV) stability in Yersinia pestis.* International Journal of Food Microbiology, 2010. **136**(3): p. 372-375.
243. Bhaduri, S., *Effect of salt and acidic pH on the stability of virulence plasmid (pYV) in Yersinia enterocolitica and expression of virulence-associated characteristics.* Food Microbiology, 2011. **28**(1): p. 171-173.
244. Gerdes, K., *Toxin-Antitoxin Modules May Regulate Synthesis of Macromolecules during Nutritional Stress.* Journal of Bacteriology, 2000. **182**(3): p. 561-572.
245. Africa, S.S., *Living Conditions of Households in South Africa, 2014/2015.* https://www.statssa.gov.za/publications/P0310/P03102014.pdf, 2017.
246. Africa, S.S., *General Household Survey.* https://www.statssa.gov.za/publications/P0318/P03182020.pdf, 2021.
247. Weill, F.-X., et al., *Genomic insights into the 2016–2017 cholera epidemic in Yemen.* Nature, 2019. **565**(7738): p. 230-233.
248. Park, S.E., et al., *The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa.* Nature Communications, 2018. **9**(1): p. 5094.
249. Van den Bossche, A., et al., *Outbreak of Central American born Shigella sonnei in two youth camps in Belgium in the summer of 2019.* European Journal of Clinical Microbiology & Infectious Diseases, 2021. **40**(7): p. 1573-1577.
250. Mikhail, A.F.W., et al., *Utility of whole-genome sequencing during an investigation of multiple foodborne outbreaks of Shigella sonnei.* Epidemiology and Infection, 2021. **149**: p. e71.
251. Baker, K.S., et al., *Travel- and community-based transmission of multidrug-resistant Shigella sonnei lineage among international Orthodox Jewish communities.* Emerging Infectious Diseases, 2016. **22**(9): p. 1545-1553.
252. Rew, V., et al., *Whole-genome sequencing revealed concurrent outbreaks of shigellosis in the English orthodox Jewish community caused by multiple importations of Shigella sonnei from Israel.* Microbial Genomics, 2018. **4**(3).
253. Terry, L.M., et al., *Antimicrobial resistance profiles of Shigella dysenteriae isolated from travellers returning to the UK, 2004–2017.* Journal of Medical Microbiology, 2018. **67**(8): p. 1022-1030.

254. Bottieau, E., et al., *Epidemiology and outcome of shigella, salmonella and campylobacter infections in travellers returning from the tropics with fever and diarrhoea.* Acta Clinica Belgica, 2011. **66**(3): p. 191-195.
255. Ekdahl, K. and Y. Andersson, *The epidemiology of travel-associated shigellosis-regional risks, seasonality and serogroups.* Journal of Infection, 2005. **51**(3): p. 222-229.
256. Toro, C., et al., *Shigellosis in subjects with traveler's diarrhea versus domestically acquired diarrhea: Implications for antimicrobial therapy and human immunodeficiency virus surveillance.* American Journal of Tropical Medicine and Hygiene, 2015. **93**(3): p. 491-496.
257. Lane, C.R., et al., *Travel Destinations and Sexual Behavior as Indicators of Antibiotic Resistant Shigella Strains - Victoria, Australia.* Clinical Infectious Diseases, 2015. **62**(6): p. 722-729.
258. Trépanier, S., et al., *Travel-related shigellosis in Quebec, Canada: An analysis of risk factors.* Journal of Travel Medicine, 2014. **21**(5): p. 304-309.
259. McGuire, E., et al., *Shigellosis in adults: A retrospective study of clinical and epidemiological features in East London.* International Journal of STD & AIDS, 2019. **30**(14): p. 1373-1381.
260. 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, 2006. **134**(6): p. 1231-6.
261. Ingle, D.J., et al., *Co-circulation of Multidrug-resistant Shigella Among Men Who Have Sex With Men in Australia.* Clinical Infectious Diseases, 2019. **69**(9): p. 1535-1544.
262. Fischer, N., et al., *Genomic epidemiology of persistently circulating MDR Shigella sonnei strains associated with men who have sex with men (MSM) in Belgium (2013–19).* Journal of Antimicrobial Chemotherapy, 2021. **77**(1): p. 89-97.
263. Moreno-Mingorance, A., et al., *Circulation of multi-drug-resistant Shigella sonnei and Shigella flexneri among men who have sex with men in Barcelona, Spain, 2015–2019.* International Journal of Antimicrobial Agents, 2021. **58**(3): p. 106378.

Supplementary

Supplementary materials table of contents

Supplementary Tables

Table 1. Malawian isolates ID and accession number (Chapter 3).

Isolate	Accession
22204_7#74	ERR2525592
22204_7#76	ERR2525594
22204_7#77	ERR2525595
22204_7#78	ERR2525596
22204_7#79	ERR2525597
22204_7#80	ERR2525598
22204_7#81	ERR2525599
22204_7#82	ERR2525600

Table 2. Isolates excluded from initial South African sample set (Chapters 4 and 5) and reason for exclusion.

Isolate	Sample received	Re-sequenced	Reason for exclusion from final dataset
FD01874177	No	No	
FD01874178	No	No	
FD01874584	No	No	
FD01874586	No	No	
FD01874602	No	No	
FD01874615	No	No	
FD01874622	No	No	
FD01874639	No	No	
FD01874665	No	No	
FD01874674	No	No	
FD01874690	No	No	
FD01874706	No	No	
FD01874765	No	No	
FD01876625	No	No	
FD01876634	No	No	
FD01876657	No	No	
FD01876667	No	No	
FD01876693	No	No	
FD01873910	Yes	No	<20x mapping coverage
FD01873912	Yes	No	<20x mapping coverage
FD01873924	Yes	No	<20x mapping coverage
FD01873989	Yes	No	<20x mapping coverage
FD01874112	Yes	No	<20x mapping coverage
FD01874173	Yes	Yes	<20x mapping coverage
FD01874759	Yes	No	<20x mapping coverage
FD01876608	Yes	No	<20x mapping coverage
FD01876694	Yes	No	<20x mapping coverage
FD01872862	Yes	Yes	GC content curve showed contamination
FD01872864	Yes	No	GC content curve showed contamination
FD01872886	Yes	Yes	GC content curve showed contamination
FD01874613	Yes	Yes	GC content curve showed contamination
FD01876616	Yes	Yes	GC content curve showed contamination
FD01874153	Yes	Yes	Shigatyper prediction: low levels, possible contamination
FD01873960	Yes	Yes	Poor per base sequence content and <20x mapping coverage
FD01873997	Yes	Yes	Poor per base sequence content and <20x mapping coverage
FD01874122	Yes	Yes	Poor per base sequence content and <20x mapping coverage
FD01874154	Yes	Yes	Poor per base sequence content and <20x mapping coverage
FD01876614	Yes	Yes	Poor per base sequence content
FD01876659	Yes	Yes	Poor per base sequence content and <20x mapping coverage

FD01872890	Yes	Yes	Shigatyper prediction: *S. flexneri* 3a, distant in ML phylogeny
FD01872923	Yes	Yes	Shigatyper prediction: *S. flexneri* 3a, distant in ML phylogeny
FD01872937	Yes	No	Shigatyper prediction: *E. coli*, distant in ML phylogeny
FD01873981	Yes	Yes	Shigatyper prediction: multiple wzx, likely contamination
FD01874188	Yes	Yes	Shigatyper prediction: no wzx, likely not *E. coli* or *Shigella*, distant in ML phylogeny
FD01874695	Yes	No	Shigatyper prediction: *E. coli*, distant in ML phylogeny
FD01874709	Yes	No	Shigatyper prediction: *E. coli*, distant in ML phylogeny
FD01874735	Yes	No	Shigatyper prediction: *E. coli*, distant in ML phylogeny
FD01876646	Yes	No	Shigatyper prediction: no wzx, likely not *E. coli* or *Shigella*, distant in ML phylogeny

Table 3. Isolates names, accession number in the European Nucleotide Archive, serotype and isolation date for all South African isolates included in Chapters 4 and 5.

Isolate	Accession number	Serotype	Isolation date	Re-sequenced
FD01872847	ERS12564449	*Shigella flexneri* 2a	17/7/2012	Yes
FD01872848	ERS12564450	*Shigella flexneri* 2a	30/9/2012	No
FD01872849	ERS12564451	*Shigella flexneri* 2a	14/12/2012	No
FD01872850	ERS12564452	*Shigella flexneri* 2a	5/2/2013	Yes
FD01872851	ERS12564453	*Shigella flexneri* 2a	7/5/2013	Yes
FD01872852	ERS12564454	*Shigella flexneri* 2a	1/8/2013	Yes
FD01872853	ERS12564455	*Shigella flexneri* 2a	21/10/2013	Yes
FD01872854	ERS12564456	*Shigella flexneri* 2a	9/12/2013	Yes
FD01872855	ERS12564457	*Shigella flexneri* 2a	8/7/2012	No
FD01872856	ERS12564458	*Shigella flexneri* 2a	15/10/2012	No
FD01872857	ERS12564459	*Shigella flexneri* 2a	24/12/2012	No
FD01872858	ERS12564460	*Shigella flexneri* 2a	3/3/2013	Yes
FD01872859	ERS12564461	*Shigella flexneri* 2a	20/4/2013	No
FD01872860	ERS12564462	*Shigella flexneri* 2a	6/8/2013	Yes
FD01872861	ERS12564463	*Shigella flexneri* 2a	23/10/2013	Yes
FD01872863	ERS12564464	*Shigella flexneri* 2a	1/8/2012	No
FD01872865	ERS12564465	*Shigella flexneri* 2a	29/12/2012	No
FD01872866	ERS12564466	*Shigella flexneri* 2a	27/2/2013	Yes
FD01872867	ERS12564467	*Shigella flexneri* 2a	23/5/2013	Yes
FD01872868	ERS12564468	*Shigella flexneri* 2a	19/8/2013	Yes
FD01872869	ERS12564469	*Shigella flexneri* 2a	6/11/2013	Yes
FD01872870	ERS12564470	*Shigella flexneri* 2a	28/12/2013	Yes
FD01872871	ERS12564471	*Shigella flexneri* 2a	10/8/2012	No
FD01872872	ERS12564472	*Shigella flexneri* 2a	23/10/2012	No
FD01872873	ERS12564473	*Shigella flexneri* 2a	4/1/2013	No
FD01872874	ERS12564474	*Shigella flexneri* 2a	26/3/2013	Yes
FD01872875	ERS12564475	*Shigella flexneri* 2a	24/5/2013	Yes
FD01872876	ERS12564476	*Shigella flexneri* 2a	29/8/2013	Yes
FD01872877	ERS12564477	*Shigella flexneri* 2a	7/11/2013	Yes
FD01872878	ERS12564478	*Shigella flexneri* 2a	23/12/2013	Yes
FD01872879	ERS12564479	*Shigella flexneri* 2a	31/7/2012	No
FD01872880	ERS12564480	*Shigella flexneri* 2a	1/11/2012	No
FD01872881	ERS12564481	*Shigella flexneri* 2a	2/1/2013	No
FD01872882	ERS12564482	*Shigella flexneri* 2a	25/3/2013	Yes
FD01872883	ERS12564483	*Shigella flexneri* 2a	30/5/2013	Yes
FD01872884	ERS12564484	*Shigella flexneri* 2a	30/8/2013	Yes
FD01872885	ERS12564485	*Shigella flexneri* 2a	11/11/2013	No
FD01872887	ERS12564486	*Shigella flexneri* 2a	23/8/2012	No
FD01872888	ERS12564487	*Shigella flexneri* 2a	2/11/2012	No
FD01872889	ERS12564488	*Shigella flexneri* 2a	22/1/2013	No
FD01872891	ERS12564489	*Shigella flexneri* 2a	10/6/2013	Yes
FD01872892	ERS12564490	*Shigella flexneri* 2a	8/9/2013	Yes
FD01872893	ERS12564491	*Shigella flexneri* 2a	13/11/2013	No
FD01872894	ERS12564492	*Shigella flexneri* 2a	2/1/2014	No
FD01872895	ERS12564493	*Shigella flexneri* 2a	2/9/2012	No

FD01872896	ERS12564494	*Shigella flexneri* 2a	14/11/2012	No
FD01872897	ERS12564495	*Shigella flexneri* 2a	13/2/2013	Yes
FD01872898	ERS12564496	*Shigella flexneri* 2a	4/4/2013	Yes
FD01872899	ERS12564497	*Shigella flexneri* 2a	14/6/2013	Yes
FD01872900	ERS12564498	*Shigella flexneri* 2a	6/9/2013	Yes
FD01872901	ERS12564499	*Shigella flexneri* 2a	30/10/2013	Yes
FD01872902	ERS12564500	*Shigella flexneri* 2a	7/1/2014	Yes
FD01872903	ERS12564501	*Shigella flexneri* 2a	8/9/2012	Yes
FD01872904	ERS12564502	*Shigella flexneri* 2a	17/11/2012	No
FD01872905	ERS12564503	*Shigella flexneri* 2a	29/1/2013	No
FD01872906	ERS12564504	*Shigella flexneri* 2a	16/4/2013	Yes
FD01872907	ERS12564505	*Shigella flexneri* 2a	29/6/2013	Yes
FD01872908	ERS12564506	*Shigella flexneri* 2a	12/9/2013	Yes
FD01872909	ERS12564507	*Shigella flexneri* 2a	15/11/2013	No
FD01872910	ERS12564508	*Shigella flexneri* 2a	16/1/2014	No
FD01872911	ERS12564509	*Shigella flexneri* 2a	21/9/2012	No
FD01872912	ERS12564510	*Shigella flexneri* 2a	29/11/2012	No
FD01872913	ERS12564511	*Shigella flexneri* 2a	11/2/2013	No
FD01872914	ERS12564512	*Shigella flexneri* 2a	25/4/2013	Yes
FD01872915	ERS12564513	*Shigella flexneri* 2a	18/6/2013	Yes
FD01872916	ERS12564514	*Shigella flexneri* 2a	21/9/2013	Yes
FD01872917	ERS12564515	*Shigella flexneri* 2a	24/11/2013	Yes
FD01872918	ERS12564516	*Shigella flexneri* 2a	13/1/2014	Yes
FD01872919	ERS12564517	*Shigella flexneri* 2a	9/9/2012	No
FD01872920	ERS12564518	*Shigella flexneri* 2a	28/11/2012	No
FD01872921	ERS12564519	*Shigella flexneri* 2a	11/11/2013	No
FD01872922	ERS12564520	*Shigella flexneri* 2a	5/5/2013	Yes
FD01872924	ERS12564521	*Shigella flexneri* 2a	13/9/2013	Yes
FD01872925	ERS12564522	*Shigella flexneri* 2a	25/11/2013	Yes
FD01872926	ERS12564523	*Shigella flexneri* 2a	20/1/2014	Yes
FD01872927	ERS12564524	*Shigella flexneri* 2a	29/9/2012	No
FD01872928	ERS12564525	*Shigella flexneri* 2a	28/11/2012	No
FD01872929	ERS12564526	*Shigella flexneri* 2a	12/1/2013	Yes
FD01872930	ERS12564527	*Shigella flexneri* 2a	6/5/2013	Yes
FD01872931	ERS12564528	*Shigella flexneri* 2a	19/7/2013	Yes
FD01872932	ERS12564529	*Shigella flexneri* 2a	15/9/2013	No
FD01872933	ERS12564530	*Shigella flexneri* 2a	3/12/2013	Yes
FD01872934	ERS12564531	*Shigella flexneri* 2a	23/1/2014	No
FD01872935	ERS12564532	*Shigella flexneri* 2a	26/9/2012	Yes
FD01872936	ERS12564533	*Shigella flexneri* 2a	11/12/2012	No
FD01872938	ERS12564534	*Shigella flexneri* 2a	9/5/2013	Yes
FD01872939	ERS12564535	*Shigella flexneri* 2a	22/7/2013	Yes
FD01872940	ERS12564536	*Shigella flexneri* 2a	11/10/2013	No
FD01872941	ERS12564537	*Shigella flexneri* 2a	7/12/2013	Yes
FD01872942	ERS12564538	*Shigella flexneri* 2a	25/1/2014	Yes
FD01873906	ERS12564539	*Shigella sonnei*	26/5/2011	No
FD01873907	ERS12564540	*Shigella sonnei*	15/10/2011	No
FD01873908	ERS12564541	*Shigella sonnei*	16/1/2012	No

FD01873909	ERS12564542	*Shigella sonnei*	21/3/2012	Yes
FD01873911	ERS12564543	*Shigella sonnei*	5/7/2012	No
FD01873914	ERS12564544	*Shigella sonnei*	20/6/2011	No
FD01873915	ERS12564545	*Shigella sonnei*	20/10/2011	No
FD01873916	ERS12564546	*Shigella sonnei*	16/3/2012	No
FD01873917	ERS12564547	*Shigella sonnei*	24/3/2012	No
FD01873918	ERS12564548	*Shigella sonnei*	19/4/2012	No
FD01873919	ERS12564549	*Shigella sonnei*	19/7/2012	No
FD01873920	ERS12564550	*Shigella sonnei*	24/10/2012	No
FD01873922	ERS12564551	*Shigella sonnei*	10/12/2011	No
FD01873923	ERS12564552	*Shigella sonnei*	3/11/2011	No
FD01873925	ERS12564553	*Shigella sonnei*	29/3/2012	No
FD01873926	ERS12564554	*Shigella sonnei*	16/4/2012	No
FD01873927	ERS12564555	*Shigella sonnei*	1/8/2012	Yes
FD01873928	ERS12564556	*Shigella sonnei*	29/10/2012	No
FD01873930	ERS12564557	*Shigella sonnei*	28/6/2011	No
FD01873931	ERS12564558	*Shigella sonnei*	11/11/2011	No
FD01873932	ERS12564559	*Shigella sonnei*	1/2/2012	No
FD01873933	ERS12564560	*Shigella sonnei*	29/3/2012	No
FD01873934	ERS12564561	*Shigella sonnei*	23/5/2012	No
FD01873935	ERS12564562	*Shigella sonnei*	29/7/2012	No
FD01873936	ERS12564563	*Shigella sonnei*	10/11/2012	No
FD01873938	ERS12564564	*Shigella sonnei*	16/7/2011	No
FD01873939	ERS12564565	*Shigella sonnei*	10/12/2011	No
FD01873940	ERS12564566	*Shigella sonnei*	6/2/2012	No
FD01873941	ERS12564567	*Shigella sonnei*	5/4/2012	No
FD01873942	ERS12564568	*Shigella sonnei*	24/5/2012	Yes
FD01873943	ERS12564569	*Shigella sonnei*	8/8/2012	No
FD01873944	ERS12564570	*Shigella sonnei*	17/11/2012	No
FD01873946	ERS12564571	*Shigella sonnei*	30/9/2011	No
FD01873947	ERS12564572	*Shigella sonnei*	27/12/2011	No
FD01873948	ERS12564573	*Shigella sonnei*	5/1/2012	No
FD01873949	ERS12564574	*Shigella sonnei*	3/4/2012	Yes
FD01873950	ERS12564575	*Shigella sonnei*	20/5/2012	No
FD01873951	ERS12564576	*Shigella sonnei*	18/8/2012	No
FD01873952	ERS12564577	*Shigella sonnei*	24/11/2012	No
FD01873954	ERS12564578	*Shigella sonnei*	6/8/2011	No
FD01873955	ERS12564579	*Shigella sonnei*	10/12/2011	No
FD01873956	ERS12564580	*Shigella sonnei*	13/2/2012	No
FD01873957	ERS12564581	*Shigella sonnei*	11/4/2012	No
FD01873958	ERS12564582	*Shigella sonnei*	7/6/2012	No
FD01873959	ERS12564583	*Shigella sonnei*	30/8/2012	No
FD01873962	ERS12564584	*Shigella sonnei*	6/7/2011	No
FD01873963	ERS12564585	*Shigella sonnei*	12/12/2011	No
FD01873964	ERS12564586	*Shigella sonnei*	28/1/2012	Yes
FD01873965	ERS12564587	*Shigella sonnei*	17/4/2012	Yes
FD01873966	ERS12564588	*Shigella sonnei*	11/6/2012	Yes
FD01873967	ERS12564589	*Shigella sonnei*	10/9/2012	Yes

FD01873968	ERS12564590	*Shigella sonnei*	5/12/2012	No
FD01873970	ERS12564591	*Shigella sonnei*	22/8/2011	No
FD01873971	ERS12564592	*Shigella sonnei*	9/12/2011	No
FD01873972	ERS12564593	*Shigella sonnei*	3/3/2012	Yes
FD01873973	ERS12564594	*Shigella sonnei*	26/4/2012	No
FD01873974	ERS12564595	*Shigella sonnei*	13/6/2012	Yes
FD01873975	ERS12564596	*Shigella sonnei*	20/9/2012	No
FD01873976	ERS12564597	*Shigella sonnei*	14/11/2012	No
FD01873978	ERS12564598	*Shigella sonnei*	26/8/2011	No
FD01873979	ERS12564599	*Shigella sonnei*	28/12/2011	No
FD01873980	ERS12564600	*Shigella sonnei*	9/3/2012	No
FD01873982	ERS12564601	*Shigella sonnei*	18/6/2012	No
FD01873983	ERS12564602	*Shigella sonnei*	25/9/2012	No
FD01873986	ERS12564603	*Shigella sonnei*	5/9/2011	No
FD01873987	ERS12564604	*Shigella sonnei*	17/3/2012	No
FD01873988	ERS12564605	*Shigella sonnei*	9/3/2012	No
FD01873990	ERS12564606	*Shigella sonnei*	19/6/2012	No
FD01873991	ERS12564607	*Shigella sonnei*	26/9/2012	No
FD01873994	ERS12564608	*Shigella sonnei*	6/10/2011	Yes
FD01873995	ERS12564609	*Shigella sonnei*	6/1/2012	No
FD01873996	ERS12564610	*Shigella sonnei*	14/3/2012	No
FD01873998	ERS12564611	*Shigella sonnei*	8/7/2012	No
FD01873999	ERS12564612	*Shigella sonnei*	18/8/2012	No
FD01874098	ERS12564613	*Shigella sonnei*	4/6/2013	No
FD01874099	ERS12564614	*Shigella sonnei*	27/10/2013	No
FD01874100	ERS12564615	*Shigella sonnei*	10/1/2014	No
FD01874101	ERS12564616	*Shigella sonnei*	20/2/2014	No
FD01874102	ERS12564617	*Shigella sonnei*	18/3/2014	No
FD01874103	ERS12564618	*Shigella sonnei*	30/4/2014	No
FD01874104	ERS12564619	*Shigella sonnei*	10/3/2011	No
FD01874105	ERS12564620	*Shigella sonnei*	26/3/2011	No
FD01874106	ERS12564621	*Shigella sonnei*	8/6/2013	No
FD01874107	ERS12564622	*Shigella sonnei*	10/11/2013	No
FD01874108	ERS12564623	*Shigella sonnei*	14/1/2014	Yes
FD01874109	ERS12564624	*Shigella sonnei*	5/2/2014	Yes
FD01874110	ERS12564625	*Shigella sonnei*	17/3/2014	No
FD01874111	ERS12564626	*Shigella sonnei*	12/5/2014	No
FD01874113	ERS12564627	*Shigella sonnei*	4/4/2011	No
FD01874114	ERS12564628	*Shigella sonnei*	21/6/2013	Yes
FD01874115	ERS12564629	*Shigella sonnei*	23/11/2013	No
FD01874116	ERS12564630	*Shigella sonnei*	3/4/2014	Yes
FD01874117	ERS12564631	*Shigella sonnei*	31/1/2014	No
FD01874118	ERS12564632	*Shigella sonnei*	25/3/2014	Yes
FD01874119	ERS12564633	*Shigella sonnei*	12/5/2014	No
FD01874120	ERS12564634	*Shigella sonnei*	23/2/2011	No
FD01874121	ERS12564635	*Shigella sonnei*	30/3/2011	No
FD01874123	ERS12564636	*Shigella sonnei*	4/12/2013	No
FD01874124	ERS12564637	*Shigella sonnei*	15/1/2014	Yes

FD01874125	ERS12564638	*Shigella sonnei*	24/2/2014	Yes
FD01874126	ERS12564639	*Shigella sonnei*	23/3/2014	Yes
FD01874127	ERS12564640	*Shigella sonnei*	23/5/2014	No
FD01874128	ERS12564641	*Shigella sonnei*	22/2/2011	No
FD01874129	ERS12564642	*Shigella sonnei*	10/4/2011	No
FD01874130	ERS12564643	*Shigella sonnei*	1/7/2013	Yes
FD01874131	ERS12564644	*Shigella sonnei*	18/12/2013	Yes
FD01874132	ERS12564645	*Shigella sonnei*	18/1/2014	Yes
FD01874133	ERS12564646	*Shigella sonnei*	24/2/2014	No
FD01874134	ERS12564647	*Shigella sonnei*	19/3/2014	Yes
FD01874135	ERS12564648	*Shigella sonnei*	2/6/2014	No
FD01874136	ERS12564649	*Shigella sonnei*	1/3/2011	No
FD01874137	ERS12564650	*Shigella sonnei*	14/4/2011	Yes
FD01874138	ERS12564651	*Shigella sonnei*	17/7/2013	Yes
FD01874139	ERS12564652	*Shigella sonnei*	9/12/2013	Yes
FD01874140	ERS12564653	*Shigella sonnei*	23/1/2014	Yes
FD01874141	ERS12564654	*Shigella sonnei*	25/2/2014	No
FD01874142	ERS12564655	*Shigella sonnei*	14/3/2014	No
FD01874143	ERS12564656	*Shigella sonnei*	10/1/2011	No
FD01874144	ERS12564657	*Shigella sonnei*	5/3/2011	No
FD01874145	ERS12564658	*Shigella sonnei*	19/4/2011	No
FD01874146	ERS12564659	*Shigella sonnei*	21/7/2013	No
FD01874147	ERS12564660	*Shigella sonnei*	2/1/2014	No
FD01874148	ERS12564661	*Shigella sonnei*	23/1/2014	No
FD01874149	ERS12564662	*Shigella sonnei*	3/3/2014	No
FD01874150	ERS12564663	*Shigella sonnei*	29/3/2014	Yes
FD01874151	ERS12564664	*Shigella sonnei*	11/1/2011	No
FD01874152	ERS12564665	*Shigella sonnei*	10/3/2011	No
FD01874155	ERS12564666	*Shigella sonnei*	4/1/2014	Yes
FD01874156	ERS12564667	*Shigella sonnei*	29/1/2014	No
FD01874157	ERS12564668	*Shigella sonnei*	6/3/2014	Yes
FD01874158	ERS12564669	*Shigella sonnei*	7/4/2014	Yes
FD01874159	ERS12564670	*Shigella sonnei*	6/1/2011	Yes
FD01874160	ERS12564671	*Shigella sonnei*	3/3/2011	No
FD01874161	ERS12564672	*Shigella sonnei*	25/4/2011	Yes
FD01874162	ERS12564673	*Shigella sonnei*	8/9/2013	No
FD01874163	ERS12564674	*Shigella sonnei*	1/1/2014	No
FD01874164	ERS12564675	*Shigella sonnei*	4/2/2014	Yes
FD01874165	ERS12564676	*Shigella sonnei*	2/5/2014	No
FD01874166	ERS12564677	*Shigella sonnei*	13/3/2014	No
FD01874167	ERS12564678	*Shigella sonnei*	28/1/2011	Yes
FD01874168	ERS12564679	*Shigella sonnei*	11/3/2011	No
FD01874169	ERS12564680	*Shigella sonnei*	4/5/2011	No
FD01874170	ERS12564681	*Shigella sonnei*	8/8/2013	No
FD01874171	ERS12564682	*Shigella sonnei*	11/1/2014	Yes
FD01874172	ERS12564683	*Shigella sonnei*	11/1/2014	Yes
FD01874174	ERS12564684	*Shigella sonnei*	16/4/2014	No
FD01874175	ERS12564685	*Shigella sonnei*	2/2/2011	No

FD01874176	ERS12564686	*Shigella sonnei*	14/6/2011	No
FD01874179	ERS12564687	*Shigella sonnei*	7/1/2014	No
FD01874180	ERS12564688	*Shigella sonnei*	12/2/2014	Yes
FD01874181	ERS12564689	*Shigella sonnei*	10/3/2014	Yes
FD01874182	ERS12564690	*Shigella sonnei*	17/4/2014	Yes
FD01874183	ERS12564691	*Shigella sonnei*	4/2/2011	No
FD01874184	ERS12564692	*Shigella sonnei*	19/3/2011	Yes
FD01874185	ERS12564693	*Shigella sonnei*	9/7/2011	Yes
FD01874186	ERS12564694	*Shigella sonnei*	18/10/2013	No
FD01874187	ERS12564695	*Shigella sonnei*	16/1/2014	No
FD01874189	ERS12564696	*Shigella sonnei*	13/3/2014	No
FD01874190	ERS12564697	*Shigella sonnei*	16/4/2014	No
FD01874191	ERS12564698	*Shigella sonnei*	29/1/2011	No
FD01874192	ERS12564699	*Shigella sonnei*	24/3/2011	No
FD01874193	ERS12564700	*Shigella sonnei*	2/6/2011	Yes
FD01874579	ERS12564701	*Shigella flexneri* 2a	25/1/2014	No
FD01874580	ERS12564702	*Shigella flexneri* 2a	7/4/2014	No
FD01874581	ERS12564703	*Shigella flexneri* 2a	8/6/2014	No
FD01874582	ERS12564704	*Shigella flexneri* 2a	5/11/2014	No
FD01874583	ERS12564705	*Shigella flexneri* 2a	19/2/2015	No
FD01874585	ERS12564706	*Shigella flexneri* 2a	24/7/2015	No
FD01874587	ERS12564707	*Shigella flexneri* 2a	16/1/2014	No
FD01874588	ERS12564708	*Shigella flexneri* 2a	8/4/2014	No
FD01874589	ERS12564709	*Shigella flexneri* 2a	20/6/2014	Yes
FD01874590	ERS12564710	*Shigella flexneri* 2a	7/11/2014	No
FD01874591	ERS12564711	*Shigella flexneri* 2a	6/3/2015	No
FD01874592	ERS12564712	*Shigella flexneri* 2a	28/3/2015	No
FD01874593	ERS12564713	*Shigella flexneri* 2a	5/8/2015	No
FD01874594	ERS12564714	*Shigella flexneri* 2a	23/11/2015	No
FD01874595	ERS12564715	*Shigella flexneri* 2a	10/2/2014	No
FD01874596	ERS12564716	*Shigella flexneri* 2a	23/5/2014	No
FD01874597	ERS12564717	*Shigella flexneri* 2a	17/7/2014	No
FD01874598	ERS12564718	*Shigella flexneri* 2a	29/10/2014	No
FD01874599	ERS12564719	*Shigella flexneri* 2a	5/3/2015	No
FD01874600	ERS12564720	*Shigella flexneri* 2a	19/5/2015	No
FD01874601	ERS12564721	*Shigella flexneri* 2a	14/8/2015	No
FD01874603	ERS12564722	*Shigella flexneri* 2a	5/2/2014	No
FD01874604	ERS12564723	*Shigella flexneri* 2a	15/4/2014	No
FD01874605	ERS12564724	*Shigella flexneri* 2a	11/8/2014	Yes
FD01874606	ERS12564725	*Shigella flexneri* 2a	1/12/2014	No
FD01874607	ERS12564726	*Shigella flexneri* 2a	2/3/2015	No
FD01874608	ERS12564727	*Shigella flexneri* 2a	24/5/2015	No
FD01874609	ERS12564728	*Shigella flexneri* 2a	21/8/2015	No
FD01874610	ERS12564729	*Shigella flexneri* 2a	12/11/2015	No
FD01874611	ERS12564730	*Shigella flexneri* 2a	25/2/2014	No
FD01874612	ERS12564731	*Shigella flexneri* 2a	9/5/2014	No
FD01874614	ERS12564732	*Shigella flexneri* 2a	17/12/2014	No
FD01874616	ERS12564733	*Shigella flexneri* 2a	15/5/2015	No

FD01874617	ERS12564734	*Shigella flexneri* 2a	3/9/2015	No
FD01874618	ERS12564735	*Shigella flexneri* 2a	15/12/2015	No
FD01874619	ERS12564736	*Shigella flexneri* 2a	27/2/2014	No
FD01874620	ERS12564737	*Shigella flexneri* 2a	22/5/2014	No
FD01874621	ERS12564738	*Shigella flexneri* 2a	26/8/2014	No
FD01874623	ERS12564739	*Shigella flexneri* 2a	2/4/2015	No
FD01874624	ERS12564740	*Shigella flexneri* 2a	31/5/2015	No
FD01874625	ERS12564741	*Shigella flexneri* 2a	8/9/2015	No
FD01874626	ERS12564742	*Shigella flexneri* 2a	21/11/2015	No
FD01874627	ERS12564743	*Shigella flexneri* 2a	26/2/2014	No
FD01874628	ERS12564744	*Shigella flexneri* 2a	9/5/2014	No
FD01874629	ERS12564745	*Shigella flexneri* 2a	8/9/2014	No
FD01874630	ERS12564746	*Shigella flexneri* 2a	20/1/2015	No
FD01874631	ERS12564747	*Shigella flexneri* 2a	7/4/2015	No
FD01874632	ERS12564748	*Shigella flexneri* 2a	25/6/2015	No
FD01874633	ERS12564749	*Shigella flexneri* 2a	14/9/2015	No
FD01874634	ERS12564750	*Shigella flexneri* 2a	17/12/2015	Yes
FD01874635	ERS12564751	*Shigella flexneri* 2a	12/3/2014	No
FD01874636	ERS12564752	*Shigella flexneri* 2a	9/6/2014	No
FD01874637	ERS12564753	*Shigella flexneri* 2a	18/9/2014	Yes
FD01874638	ERS12564754	*Shigella flexneri* 2a	22/1/2015	Yes
FD01874640	ERS12564755	*Shigella flexneri* 2a	28/6/2015	Yes
FD01874641	ERS12564756	*Shigella flexneri* 2a	3/10/2015	Yes
FD01874642	ERS12564757	*Shigella flexneri* 2a	14/11/2015	No
FD01874643	ERS12564758	*Shigella flexneri* 2a	14/3/2014	Yes
FD01874644	ERS12564759	*Shigella flexneri* 2a	12/6/2014	No
FD01874645	ERS12564760	*Shigella flexneri* 2a	17/8/2014	Yes
FD01874646	ERS12564761	*Shigella flexneri* 2a	26/1/2015	No
FD01874647	ERS12564762	*Shigella flexneri* 2a	15/4/2015	No
FD01874649	ERS12564763	*Shigella flexneri* 2a	22/6/2015	No
FD01874650	ERS12564764	*Shigella flexneri* 2a	6/10/2015	No
FD01874651	ERS12564765	*Shigella flexneri* 2a	21/12/2015	No
FD01874652	ERS12564766	*Shigella flexneri* 2a	6/3/2014	No
FD01874653	ERS12564767	*Shigella flexneri* 2a	14/3/2014	No
FD01874654	ERS12564768	*Shigella flexneri* 2a	4/9/2014	No
FD01874655	ERS12564769	*Shigella flexneri* 2a	2/2/2015	No
FD01874656	ERS12564770	*Shigella flexneri* 2a	27/4/2015	No
FD01874657	ERS12564771	*Shigella flexneri* 2a	3/7/2015	No
FD01874658	ERS12564772	*Shigella flexneri* 2a	15/10/2015	No
FD01874659	ERS12564773	*Shigella flexneri* 2a	11/11/2015	No
FD01874660	ERS12564774	*Shigella flexneri* 2a	6/3/2014	No
FD01874661	ERS12564775	*Shigella flexneri* 2a	12/6/2014	No
FD01874662	ERS12564776	*Shigella flexneri* 2a	6/10/2014	No
FD01874663	ERS12564777	*Shigella flexneri* 2a	7/2/2015	No
FD01874664	ERS12564778	*Shigella flexneri* 2a	21/4/2015	No
FD01874666	ERS12564779	*Shigella flexneri* 2a	27/10/2015	Yes
FD01874667	ERS12564780	*Shigella flexneri* 2a	14/6/2014	Yes
FD01874668	ERS12564781	*Shigella flexneri* 2a	24/3/2014	No

FD01874669	ERS12564782	*Shigella flexneri* 2a	24/6/2014	No
FD01874670	ERS12564783	*Shigella flexneri* 2a	14/10/2014	Yes
FD01874671	ERS12564784	*Shigella flexneri* 2a	3/2/2015	No
FD01874672	ERS12564785	*Shigella flexneri* 2a	4/5/2015	No
FD01874673	ERS12564786	*Shigella flexneri* 2a	23/7/2015	No
FD01874675	ERS12564787	*Shigella sonnei*	22/3/2014	Yes
FD01874676	ERS12564788	*Shigella sonnei*	22/6/2014	No
FD01874677	ERS12564789	*Shigella sonnei*	4/2/2015	No
FD01874678	ERS12564790	*Shigella sonnei*	5/5/2015	No
FD01874679	ERS12564791	*Shigella sonnei*	24/8/2015	No
FD01874680	ERS12564792	*Shigella sonnei*	7/12/2015	No
FD01874681	ERS12564793	*Shigella sonnei*	6/2/2013	No
FD01874682	ERS12564794	*Shigella sonnei*	28/2/2013	No
FD01874683	ERS12564795	*Shigella sonnei*	4/4/2013	No
FD01874684	ERS12564796	*Shigella sonnei*	28/7/2014	No
FD01874685	ERS12564797	*Shigella sonnei*	9/2/2015	No
FD01874686	ERS12564798	*Shigella sonnei*	14/5/2015	No
FD01874687	ERS12564799	*Shigella sonnei*	10/9/2015	No
FD01874688	ERS12564800	*Shigella sonnei*	10/12/2012	No
FD01874689	ERS12564801	*Shigella sonnei*	5/2/2013	No
FD01874691	ERS12564802	*Shigella sonnei*	10/4/2013	No
FD01874692	ERS12564803	*Shigella sonnei*	31/7/2014	No
FD01874693	ERS12564804	*Shigella sonnei*	13/2/2015	No
FD01874694	ERS12564805	*Shigella sonnei*	21/4/2015	No
FD01874696	ERS12564806	*Shigella sonnei*	7/12/2012	No
FD01874697	ERS12564807	*Shigella sonnei*	8/2/2013	No
FD01874698	ERS12564808	*Shigella sonnei*	24/1/2013	No
FD01874699	ERS12564809	*Shigella sonnei*	11/4/2013	No
FD01874700	ERS12564810	*Shigella sonnei*	12/8/2014	No
FD01874701	ERS12564811	*Shigella sonnei*	21/2/2015	No
FD01874702	ERS12564812	*Shigella sonnei*	23/5/2015	Yes
FD01874703	ERS12564813	*Shigella sonnei*	23/9/2015	No
FD01874704	ERS12564814	*Shigella sonnei*	30/12/2012	No
FD01874705	ERS12564815	*Shigella sonnei*	16/2/2013	No
FD01874707	ERS12564816	*Shigella sonnei*	12/4/2013	No
FD01874708	ERS12564817	*Shigella sonnei*	3/9/2014	No
FD01874710	ERS12564818	*Shigella sonnei*	3/5/2015	Yes
FD01874711	ERS12564819	*Shigella sonnei*	3/10/2015	No
FD01874712	ERS12564820	*Shigella sonnei*	28/12/2012	No
FD01874713	ERS12564821	*Shigella sonnei*	9/2/2013	No
FD01874714	ERS12564822	*Shigella sonnei*	10/3/2013	No
FD01874715	ERS12564823	*Shigella sonnei*	13/4/2013	Yes
FD01874716	ERS12564824	*Shigella sonnei*	4/9/2014	No
FD01874717	ERS12564825	*Shigella sonnei*	5/3/2015	No
FD01874718	ERS12564826	*Shigella sonnei*	27/5/2015	Yes
FD01874719	ERS12564827	*Shigella sonnei*	28/10/2015	Yes
FD01874720	ERS12564828	*Shigella sonnei*	6/1/2013	No
FD01874721	ERS12564829	*Shigella sonnei*	13/2/2013	No

FD01874722	ERS12564830	*Shigella sonnei*	18/3/2013	No
FD01874723	ERS12564831	*Shigella sonnei*	18/4/2013	No
FD01874724	ERS12564832	*Shigella sonnei*	14/9/2014	No
FD01874725	ERS12564833	*Shigella sonnei*	6/3/2015	Yes
FD01874726	ERS12564834	*Shigella sonnei*	13/6/2015	No
FD01874727	ERS12564835	*Shigella sonnei*	10/11/2015	No
FD01874728	ERS12564836	*Shigella sonnei*	9/1/2013	No
FD01874729	ERS12564837	*Shigella sonnei*	11/2/2013	No
FD01874730	ERS12564838	*Shigella sonnei*	20/3/2013	Yes
FD01874731	ERS12564839	*Shigella sonnei*	17/4/2013	No
FD01874732	ERS12564840	*Shigella sonnei*	8/9/2014	No
FD01874733	ERS12564841	*Shigella sonnei*	12/3/2015	No
FD01874734	ERS12564842	*Shigella sonnei*	27/6/2015	No
FD01874736	ERS12564843	*Shigella sonnei*	30/1/2013	No
FD01874737	ERS12564844	*Shigella sonnei*	19/2/2013	No
FD01874738	ERS12564845	*Shigella sonnei*	18/3/2013	No
FD01874739	ERS12564846	*Shigella sonnei*	3/5/2013	No
FD01874740	ERS12564847	*Shigella sonnei*	22/10/2014	No
FD01874741	ERS12564848	*Shigella sonnei*	30/3/2015	Yes
FD01874742	ERS12564849	*Shigella sonnei*	17/6/2015	No
FD01874743	ERS12564850	*Shigella sonnei*	19/11/2015	No
FD01874744	ERS12564851	*Shigella sonnei*	20/1/2013	No
FD01874745	ERS12564852	*Shigella sonnei*	23/2/2013	No
FD01874746	ERS12564853	*Shigella sonnei*	23/3/2013	No
FD01874747	ERS12564854	*Shigella sonnei*	8/5/2013	No
FD01874748	ERS12564855	*Shigella sonnei*	5/11/2014	Yes
FD01874749	ERS12564856	*Shigella sonnei*	7/4/2015	No
FD01874750	ERS12564857	*Shigella sonnei*	14/6/2015	No
FD01874751	ERS12564858	*Shigella sonnei*	15/11/2015	No
FD01874752	ERS12564859	*Shigella sonnei*	26/1/2013	No
FD01874753	ERS12564860	*Shigella sonnei*	26/2/2013	No
FD01874754	ERS12564861	*Shigella sonnei*	28/3/2013	No
FD01874755	ERS12564862	*Shigella sonnei*	21/5/2013	No
FD01874756	ERS12564863	*Shigella sonnei*	19/10/2014	Yes
FD01874757	ERS12564864	*Shigella sonnei*	17/3/2015	No
FD01874758	ERS12564865	*Shigella sonnei*	25/7/2015	Yes
FD01874760	ERS12564866	*Shigella sonnei*	28/1/2013	No
FD01874761	ERS12564867	*Shigella sonnei*	26/2/2013	No
FD01874762	ERS12564868	*Shigella sonnei*	25/3/2013	No
FD01874763	ERS12564869	*Shigella sonnei*	19/5/2013	No
FD01874764	ERS12564870	*Shigella sonnei*	31/12/2014	No
FD01874766	ERS12564871	*Shigella sonnei*	1/8/2015	No
FD01874767	ERS12564872	*Shigella sonnei*	10/12/2015	No
FD01874768	ERS12564873	*Shigella sonnei*	2/2/2013	No
FD01874769	ERS12564874	*Shigella sonnei*	24/2/2013	No
FD01874770	ERS12564875	*Shigella sonnei*	25/3/2013	No
FD01874771	ERS12564876	*Shigella sonnei*	25/5/2013	Yes
FD01876599	ERS12564877	*Shigella flexneri* 2a	6/1/2011	Yes

FD01876600	ERS12564878	*Shigella flexneri* 2a	7/2/2011	No
FD01876601	ERS12564879	*Shigella flexneri* 2a	16/2/2011	No
FD01876602	ERS12564880	*Shigella flexneri* 2a	27/6/2011	No
FD01876603	ERS12564881	*Shigella flexneri* 2a	9/9/2011	Yes
FD01876604	ERS12564882	*Shigella flexneri* 2a	21/11/2011	No
FD01876606	ERS12564883	*Shigella flexneri* 2a	1/2/2012	Yes
FD01876607	ERS12564884	*Shigella flexneri* 2a	16/3/2012	No
FD01876609	ERS12564885	*Shigella flexneri* 2a	26/2/2011	No
FD01876610	ERS12564886	*Shigella flexneri* 2a	24/4/2011	No
FD01876611	ERS12564887	*Shigella flexneri* 2a	11/7/2011	No
FD01876612	ERS12564888	*Shigella flexneri* 2a	9/9/2011	No
FD01876613	ERS12564889	*Shigella flexneri* 2a	19/11/2011	Yes
FD01876615	ERS12564890	*Shigella flexneri* 2a	29/4/2012	Yes
FD01876617	ERS12564891	*Shigella flexneri* 2a	1/3/2011	No
FD01876618	ERS12564892	*Shigella flexneri* 2a	25/4/2011	No
FD01876619	ERS12564893	*Shigella flexneri* 2a	15/6/2011	No
FD01876620	ERS12564894	*Shigella flexneri* 2a	4/10/2011	No
FD01876621	ERS12564895	*Shigella flexneri* 2a	18/12/2011	Yes
FD01876622	ERS12564896	*Shigella flexneri* 2a	10/2/2012	No
FD01876623	ERS12564897	*Shigella flexneri* 2a	14/5/2012	Yes
FD01876624	ERS12564898	*Shigella flexneri* 2a	14/1/2011	No
FD01876626	ERS12564899	*Shigella flexneri* 2a	11/5/2011	Yes
FD01876627	ERS12564900	*Shigella flexneri* 2a	6/7/2011	No
FD01876628	ERS12564901	*Shigella flexneri* 2a	4/8/2011	Yes
FD01876629	ERS12564902	*Shigella flexneri* 2a	21/12/2011	Yes
FD01876630	ERS12564903	*Shigella flexneri* 2a	7/2/2012	Yes
FD01876631	ERS12564904	*Shigella flexneri* 2a	17/5/2012	Yes
FD01876632	ERS12564905	*Shigella flexneri* 2a	9/1/2011	Yes
FD01876633	ERS12564906	*Shigella flexneri* 2a	10/3/2011	Yes
FD01876635	ERS12564907	*Shigella flexneri* 2a	27/7/2011	No
FD01876636	ERS12564908	*Shigella flexneri* 2a	13/10/2011	No
FD01876637	ERS12564909	*Shigella flexneri* 2a	21/12/2011	Yes
FD01876638	ERS12564910	*Shigella flexneri* 2a	18/2/2012	Yes
FD01876639	ERS12564911	*Shigella flexneri* 2a	20/5/2012	Yes
FD01876640	ERS12564912	*Shigella flexneri* 2a	18/1/2011	Yes
FD01876641	ERS12564913	*Shigella flexneri* 2a	17/3/2011	No
FD01876642	ERS12564914	*Shigella flexneri* 2a	24/5/2011	Yes
FD01876643	ERS12564915	*Shigella flexneri* 2a	2/8/2011	Yes
FD01876644	ERS12564916	*Shigella flexneri* 2a	12/10/2011	Yes
FD01876645	ERS12564917	*Shigella flexneri* 2a	24/12/2011	No
FD01876647	ERS12564918	*Shigella flexneri* 2a	22/4/2012	No
FD01876648	ERS12564919	*Shigella flexneri* 2a	26/1/2011	Yes
FD01876649	ERS12564920	*Shigella flexneri* 2a	21/2/2011	Yes
FD01876650	ERS12564921	*Shigella flexneri* 2a	12/5/2011	No
FD01876651	ERS12564922	*Shigella flexneri* 2a	4/8/2011	No
FD01876652	ERS12564923	*Shigella flexneri* 2a	26/11/2011	No
FD01876653	ERS12564924	*Shigella flexneri* 2a	9/1/2012	No
FD01876654	ERS12564925	*Shigella flexneri* 2a	15/3/2012	No

FD01876655	ERS12564926	*Shigella flexneri* 2a	29/5/2012	Yes
FD01876656	ERS12564927	*Shigella flexneri* 2a	9/1/2011	No
FD01876658	ERS12564928	*Shigella flexneri* 2a	6/6/2011	Yes
FD01876660	ERS12564929	*Shigella flexneri* 2a	21/9/2011	Yes
FD01876661	ERS12564930	*Shigella flexneri* 2a	3/1/2012	Yes
FD01876662	ERS12564931	*Shigella flexneri* 2a	25/3/2012	Yes
FD01876663	ERS12564932	*Shigella flexneri* 2a	7/6/2012	Yes
FD01876664	ERS12564933	*Shigella flexneri* 2a	31/1/2011	Yes
FD01876665	ERS12564934	*Shigella flexneri* 2a	2/4/2011	Yes
FD01876666	ERS12564935	*Shigella flexneri* 2a	25/5/2011	No
FD01876668	ERS12564936	*Shigella flexneri* 2a	18/10/2011	No
FD01876669	ERS12564937	*Shigella flexneri* 2a	7/1/2012	Yes
FD01876670	ERS12564938	*Shigella flexneri* 2a	27/2/2012	Yes
FD01876671	ERS12564939	*Shigella flexneri* 2a	6/6/2012	Yes
FD01876672	ERS12564940	*Shigella flexneri* 2a	15/2/2011	Yes
FD01876673	ERS12564941	*Shigella flexneri* 2a	9/4/2011	No
FD01876674	ERS12564942	*Shigella flexneri* 2a	19/9/2011	Yes
FD01876675	ERS12564943	*Shigella flexneri* 2a	19/8/2011	Yes
FD01876676	ERS12564944	*Shigella flexneri* 2a	31/10/2011	Yes
FD01876677	ERS12564945	*Shigella flexneri* 2a	16/1/2012	Yes
FD01876678	ERS12564946	*Shigella flexneri* 2a	9/4/2012	No
FD01876679	ERS12564947	*Shigella flexneri* 2a	21/6/2012	Yes
FD01876680	ERS12564948	*Shigella flexneri* 2a	21/2/2011	Yes
FD01876681	ERS12564949	*Shigella flexneri* 2a	21/4/2011	No
FD01876682	ERS12564950	*Shigella flexneri* 2a	13/6/2011	Yes
FD01876683	ERS12564951	*Shigella flexneri* 2a	30/8/2011	No
FD01876684	ERS12564952	*Shigella flexneri* 2a	7/11/2011	Yes
FD01876685	ERS12564953	*Shigella flexneri* 2a	24/1/2012	Yes
FD01876686	ERS12564954	*Shigella flexneri* 2a	11/4/2012	Yes
FD01876687	ERS12564955	*Shigella flexneri* 2a	10/6/2012	Yes
FD01876688	ERS12564956	*Shigella flexneri* 2a	13/2/2011	Yes
FD01876689	ERS12564957	*Shigella flexneri* 2a	29/4/2011	No
FD01876690	ERS12564958	*Shigella flexneri* 2a	24/6/2011	No
FD01876691	ERS12564959	*Shigella flexneri* 2a	8/9/2011	No
FD01876692	ERS12564960	*Shigella flexneri* 2a	8/11/2011	No
FD01876695	ERS12564961	*Shigella flexneri* 2a	4/7/2012	Yes

Table 4. Province and district where isolate collected, degree of urbanisation of the district of collection, gender and age of patient isolate collected from, for South African isolates included in Chapters 4 and 5.
DM = district municipality, MM = metropolitan municipality.

Isolate	Province	District	Degree of urbanisation	Gender	Age (years)
FD01872847	Gauteng	Ekurhuleni MM	Densely populated area	Male	Unknown
FD01872848	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	2
FD01872849	Mpumalanga	Ehlanzeni DM	Intermediate density area	Female	Unknown
FD01872850	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	2
FD01872851	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	2
FD01872852	Western Cape	City of Cape Town MM	Densely populated area	Male	3
FD01872853	Gauteng	West Rand DM	Densely populated area	Female	Unknown
FD01872854	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Unknown	2
FD01872855	Western Cape	City of Cape Town MM	Densely populated area	Female	3
FD01872856	Western Cape	City of Cape Town MM	Densely populated area	Female	6
FD01872857	Gauteng	City of Tshwane MM	Densely populated area	Female	38
FD01872858	Free State	Mangaung MM	Densely populated area	Male	2
FD01872859	Gauteng	City of Tshwane MM	Densely populated area	Male	2
FD01872860	Western Cape	City of Cape Town MM	Densely populated area	Female	4
FD01872861	Western Cape	West Coast DM	Thinly populated area	Male	1
FD01872863	Gauteng	Unknown	Unknown	Female	Unknown
FD01872865	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01872866	Western Cape	City of Cape Town MM	Densely populated area	Male	23
FD01872867	Gauteng	City of Johannesburg MM	Densely populated area	Male	11
FD01872868	Gauteng	City of Johannesburg MM	Densely populated area	Male	4

FD01872869	Eastern Cape	Amathole DM	Thinly populated area	Male	3
FD01872870	Gauteng	Ekurhuleni MM	Densely populated area	Male	3
FD01872871	Western Cape	City of Cape Town MM	Densely populated area	Male	40
FD01872872	Western Cape	West Coast DM	Thinly populated area	Female	30
FD01872873	Western Cape	Garden Route DM	Thinly populated area	Female	57
FD01872874	Eastern Cape	Amathole DM	Thinly populated area	Female	27
FD01872875	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01872876	Gauteng	City of Tshwane MM	Densely populated area	Female	55
FD01872877	Gauteng	City of Johannesburg MM	Densely populated area	Female	6
FD01872878	Eastern Cape	Joe Gqabi DM	Thinly populated area	Male	1
FD01872879	Western Cape	City of Cape Town MM	Densely populated area	Female	0
FD01872880	Northern Cape	Frances Baard DM	Intermediate density area	Male	61
FD01872881	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	1
FD01872882	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01872883	Gauteng	City of Tshwane MM	Densely populated area	Male	62
FD01872884	Gauteng	City of Johannesburg MM	Densely populated area	Female	3
FD01872885	Gauteng	City of Johannesburg MM	Densely populated area	Male	4
FD01872887	Western Cape	Garden Route DM	Thinly populated area	Female	2
FD01872888	Eastern Cape	Amathole DM	Thinly populated area	Male	7
FD01872889	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	53
FD01872891	Free State	Mangaung MM	Densely populated area	Male	56
FD01872892	Western Cape	City of Cape Town MM	Densely populated area	Male	1

FD01872893	Gauteng	West Rand DM	Densely populated area	Male	4
FD01872894	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	1
FD01872895	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	5
FD01872896	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01872897	Gauteng	Ekurhuleni MM	Densely populated area	Female	Unknown
FD01872898	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	54
FD01872899	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	1
FD01872900	Western Cape	Central Karoo DM	Thinly populated area	Male	19
FD01872901	Eastern Cape	Amathole DM	Thinly populated area	Male	0
FD01872902	Gauteng	City of Johannesburg MM	Densely populated area	Male	6
FD01872903	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	4
FD01872904	Mpumalanga	Ehlanzeni DM	Intermediate density area	Male	2
FD01872905	Western Cape	City of Cape Town MM	Densely populated area	Male	100
FD01872906	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	8
FD01872907	Free State	Lejweleputswa DM	Intermediate density area	Female	1
FD01872908	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	0
FD01872909	Western Cape	City of Cape Town MM	Densely populated area	Male	64
FD01872910	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	35
FD01872911	Free State	Lejweleputswa DM	Intermediate density area	Female	3
FD01872912	Gauteng	Sedibeng DM	Densely populated area	Male	2
FD01872913	Free State	Mangaung MM	Densely populated area	Male	71
FD01872914	Gauteng	West Rand DM	Densely populated area	Male	1

FD01872915	KwaZulu-Natal	iLembe DM	Thinly populated area	Male	2
FD01872916	Gauteng	City of Johannesburg MM	Densely populated area	Male	4
FD01872917	KwaZulu-Natal	City of eThekwini MM	Densely populated area	UNK	1
FD01872918	Gauteng	City of Johannesburg MM	Densely populated area	Female	1
FD01872919	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	0
FD01872920	Eastern Cape	Amathole DM	Thinly populated area	Female	1
FD01872921	Free State	Lejweleputswa DM	Intermediate density area	Female	63
FD01872922	Gauteng	Sedibeng DM	Densely populated area	Male	2
FD01872924	Western Cape	City of Cape Town MM	Densely populated area	Male	2
FD01872925	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	1
FD01872926	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	36
FD01872927	Western Cape	City of Cape Town MM	Densely populated area	Male	27
FD01872928	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	0
FD01872929	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01872930	Gauteng	Ekurhuleni MM	Densely populated area	Male	Unknown
FD01872931	Gauteng	City of Tshwane MM	Densely populated area	Female	12
FD01872932	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	1
FD01872933	Western Cape	West Coast DM	Thinly populated area	Female	0
FD01872934	Mpumalanga	Gert Sibande DM	Intermediate density area	Female	0
FD01872935	Western Cape	City of Cape Town MM	Densely populated area	Female	3
FD01872936	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01872938	Gauteng	City of Johannesburg MM	Densely populated area	Female	1

FD01872939	Gauteng	City of Tshwane MM	Densely populated area	Female	38
FD01872940	Western Cape	City of Cape Town MM	Densely populated area	Female	36
FD01872941	Gauteng	Sedibeng DM	Densely populated area	Female	58
FD01872942	Gauteng	City of Johannesburg MM	Densely populated area	Female	2
FD01873906	Western Cape	City of Cape Town MM	Densely populated area	Male	0
FD01873907	Gauteng	Sedibeng DM	Densely populated area	Male	1
FD01873908	Gauteng	City of Tshwane MM	Densely populated area	Male	7
FD01873909	Gauteng	West Rand DM	Densely populated area	Male	36
FD01873911	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	7
FD01873914	Gauteng	Ekurhuleni MM	Densely populated area	Male	55
FD01873915	Western Cape	Cape Winelands DM	Intermediate density area	Female	80
FD01873916	Gauteng	Unknown	Unknown	Male	Unknown
FD01873917	Gauteng	Sedibeng DM	Densely populated area	Female	4
FD01873918	Western Cape	City of Cape Town MM	Densely populated area	Male	5
FD01873919	Eastern Cape	Amathole DM	Thinly populated area	Male	Unknown
FD01873920	Free State	Mangaung MM	Densely populated area	Female	22
FD01873922	Gauteng	City of Johannesburg MM	Densely populated area	Female	6
FD01873923	Gauteng	Ekurhuleni MM	Densely populated area	Male	72
FD01873925	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01873926	KwaZulu-Natal	Amajuba DM	Thinly populated area	Male	1
FD01873927	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	8
FD01873928	Mpumalanga	Ehlanzeni DM	Intermediate density area	Female	57
FD01873930	Gauteng	Unknown	Unknown	Male	30

FD01873931	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01873932	Gauteng	Ekurhuleni MM	Densely populated area	Male	Unknown
FD01873933	Free State	Mangaung MM	Densely populated area	Female	5
FD01873934	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	2
FD01873935	Free State	Mangaung MM	Densely populated area	Female	Unknown
FD01873936	Eastern Cape	Amathole DM	Thinly populated area	Female	12
FD01873938	Gauteng	Unknown	Unknown	Female	Unknown
FD01873939	Gauteng	City of Johannesburg MM	Densely populated area	Male	5
FD01873940	Gauteng	Ekurhuleni MM	Densely populated area	Female	Unknown
FD01873941	Gauteng	Sedibeng DM	Densely populated area	Male	3
FD01873942	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	0
FD01873943	Mpumalanga	Ehlanzeni DM	Intermediate density area	Female	4
FD01873944	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01873946	Gauteng	Sedibeng DM	Densely populated area	Female	Unknown
FD01873947	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01873948	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	6
FD01873949	Eastern Cape	Unknown	Unknown	Male	31
FD01873950	Western Cape	City of Cape Town MM	Densely populated area	Female	57
FD01873951	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01873952	Gauteng	Ekurhuleni MM	Densely populated area	Female	6
FD01873954	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01873955	Gauteng	City of Johannesburg MM	Densely populated area	Male	4
FD01873956	Gauteng	Sedibeng DM	Densely populated area	Male	Unknown
FD01873957	Gauteng	Unknown	Unknown	Female	Unknown

FD01873958	Gauteng	Unknown	Unknown	Female	Unknown
FD01873959	Gauteng	City of Johannesburg MM	Densely populated area	Male	3
FD01873962	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	Unknown
FD01873963	Western Cape	City of Cape Town MM	Densely populated area	Male	1
FD01873964	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01873965	Eastern Cape	Amathole DM	Thinly populated area	Female	4
FD01873966	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01873967	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01873968	Mpumalanga	Ehlanzeni DM	Intermediate density area	Female	49
FD01873970	Western Cape	City of Cape Town MM	Densely populated area	Male	13
FD01873971	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	22
FD01873972	Gauteng	City of Johannesburg MM	Densely populated area	Male	39
FD01873973	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01873974	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01873975	Gauteng	City of Johannesburg MM	Densely populated area	Female	8
FD01873976	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	1
FD01873978	Western Cape	Unknown	Unknown	Male	18
FD01873979	KwaZulu-Natal	Unknown	Unknown	Female	8
FD01873980	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01873982	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01873983	Gauteng	Ekurhuleni MM	Densely populated area	Female	4
FD01873986	Gauteng	City of Tshwane MM	Densely populated area	Female	Unknown
FD01873987	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown

FD01873988	Gauteng	Sedibeng DM	Densely populated area	Female	14
FD01873990	Gauteng	Ekurhuleni MM	Densely populated area	Female	Unknown
FD01873991	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01873994	Gauteng	City of Tshwane MM	Densely populated area	Female	Unknown
FD01873995	Gauteng	City of Tshwane MM	Densely populated area	Female	40
FD01873996	Western Cape	Garden Route DM	Thinly populated area	Male	7
FD01873998	Gauteng	City of Tshwane MM	Densely populated area	Male	8
FD01873999	Western Cape	City of Cape Town MM	Densely populated area	UNK	3
FD01874098	Gauteng	City of Tshwane MM	Densely populated area	Male	38
FD01874099	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874100	North West	Bojanala Platinum DM	Intermediate density area	Female	32
FD01874101	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874102	Free State	Mangaung MM	Densely populated area	Female	4
FD01874103	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	4
FD01874104	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874105	Gauteng	Unknown	Unknown	Male	Unknown
FD01874106	Gauteng	Ekurhuleni MM	Densely populated area	Female	8
FD01874107	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	63
FD01874108	Gauteng	City of Johannesburg MM	Densely populated area	Male	3
FD01874109	Western Cape	City of Cape Town MM	Densely populated area	Female	4
FD01874110	Gauteng	City of Johannesburg MM	Densely populated area	Male	12
FD01874111	Western Cape	City of Cape Town MM	Densely populated area	Male	33
FD01874113	Western Cape	City of Cape Town MM	Densely populated area	Male	1

FD01874114	Mpumalanga	Ehlanzeni DM	Intermediate density area	Male	0
FD01874115	Gauteng	City of Johannesburg MM	Densely populated area	Female	0
FD01874116	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	7
FD01874117	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Female	27
FD01874118	Gauteng	Ekurhuleni MM	Densely populated area	Male	Unknown
FD01874119	Gauteng	City of Johannesburg MM	Densely populated area	Male	14
FD01874120	Eastern Cape	Amathole DM	Thinly populated area	Female	Unknown
FD01874121	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01874123	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874124	Gauteng	City of Tshwane MM	Densely populated area	Male	2
FD01874125	Gauteng	Ekurhuleni MM	Densely populated area	Male	1
FD01874126	Free State	Lejweleputswa DM	Intermediate density area	Female	56
FD01874127	Gauteng	City of Tshwane MM	Densely populated area	Male	5
FD01874128	Eastern Cape	Amathole DM	Thinly populated area	Male	0
FD01874129	Gauteng	City of Tshwane MM	Densely populated area	Female	1
FD01874130	Gauteng	City of Tshwane MM	Densely populated area	Male	2
FD01874131	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874132	Gauteng	City of Johannesburg MM	Densely populated area	Female	59
FD01874133	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874134	Western Cape	City of Cape Town MM	Densely populated area	Male	13
FD01874135	Gauteng	Ekurhuleni MM	Densely populated area	Female	5
FD01874136	Gauteng	City of Tshwane MM	Densely populated area	Female	2

FD01874137	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874138	Gauteng	City of Johannesburg MM	Densely populated area	Female	28
FD01874139	Eastern Cape	Amathole DM	Thinly populated area	Female	27
FD01874140	Western Cape	City of Cape Town MM	Densely populated area	Male	50
FD01874141	North West	Dr Kenneth Kaunda MM	Intermediate density area	Female	1
FD01874142	Western Cape	West Coast DM	Thinly populated area	Male	2
FD01874143	Eastern Cape	Amathole DM	Thinly populated area	Female	57
FD01874144	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874145	Gauteng	Ekurhuleni MM	Densely populated area	Male	2
FD01874146	Gauteng	City of Johannesburg MM	Densely populated area	Female	39
FD01874147	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	4
FD01874148	Gauteng	City of Tshwane MM	Densely populated area	Female	3
FD01874149	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874150	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874151	Western Cape	City of Cape Town MM	Densely populated area	Male	10
FD01874152	Gauteng	City of Tshwane MM	Densely populated area	Female	21
FD01874155	Gauteng	City of Tshwane MM	Densely populated area	Female	26
FD01874156	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	1
FD01874157	Gauteng	City of Johannesburg MM	Densely populated area	Male	10
FD01874158	Gauteng	City of Johannesburg MM	Densely populated area	Male	32
FD01874159	Gauteng	Unknown	Unknown	Male	Unknown
FD01874160	Gauteng	City of Tshwane MM	Densely populated area	Male	1
FD01874161	Gauteng	City of Johannesburg MM	Densely populated area	Female	11

FD01874162	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	3
FD01874163	Gauteng	City of Johannesburg MM	Densely populated area	Female	2
FD01874164	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	4
FD01874165	Gauteng	City of Tshwane MM	Densely populated area	Male	3
FD01874166	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Female	0
FD01874167	Gauteng	Unknown	Unknown	Female	22
FD01874168	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874169	Western Cape	City of Cape Town MM	Densely populated area	Male	1
FD01874170	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	4
FD01874171	Gauteng	City of Johannesburg MM	Densely populated area	Male	4
FD01874172	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	3
FD01874174	Western Cape	City of Cape Town MM	Densely populated area	Male	34
FD01874175	Gauteng	Unknown	Unknown	Male	48
FD01874176	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	1
FD01874179	Gauteng	City of Tshwane MM	Densely populated area	Female	57
FD01874180	Western Cape	City of Cape Town MM	Densely populated area	Male	4
FD01874181	Gauteng	Ekurhuleni MM	Densely populated area	Male	5
FD01874182	Gauteng	City of Tshwane MM	Densely populated area	Female	23
FD01874183	Northern Cape	Frances Baard DM	Intermediate density area	Male	Unknown
FD01874184	Gauteng	West Rand DM	Densely populated area	Female	Unknown
FD01874185	Gauteng	City of Johannesburg MM	Densely populated area	Male	2
FD01874186	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01874187	Gauteng	City of Johannesburg MM	Densely populated area	Male	38

FD01874189	Gauteng	West Rand DM	Densely populated area	Female	Unknown
FD01874190	Eastern Cape	Amathole DM	Thinly populated area	Male	1
FD01874191	Gauteng	City of Tshwane MM	Densely populated area	Female	3
FD01874192	Gauteng	Unknown	Unknown	Male	0
FD01874193	Western Cape	City of Cape Town MM	Densely populated area	Female	35
FD01874579	Western Cape	City of Cape Town MM	Densely populated area	Female	4
FD01874580	North West	Bojanala Platinum DM	Intermediate density area	Female	2
FD01874581	Western Cape	City of Cape Town MM	Densely populated area	Male	Unknown
FD01874582	Western Cape	City of Cape Town MM	Densely populated area	Male	5
FD01874583	Western Cape	City of Cape Town MM	Densely populated area	Female	4
FD01874585	Free State	Mangaung MM	Densely populated area	Male	3
FD01874587	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Male	8
FD01874588	Western Cape	City of Cape Town MM	Densely populated area	Male	1
FD01874589	Eastern Cape	Amathole DM	Thinly populated area	Female	1
FD01874590	Gauteng	City of Tshwane MM	Densely populated area	Male	1
FD01874591	Western Cape	City of Cape Town MM	Densely populated area	Female	35
FD01874592	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	17
FD01874593	Eastern Cape	Amathole DM	Thinly populated area	Male	39
FD01874594	Western Cape	City of Cape Town MM	Densely populated area	Male	1
FD01874595	Eastern Cape	Amathole DM	Thinly populated area	Female	1
FD01874596	Gauteng	City of Tshwane MM	Densely populated area	Male	5
FD01874597	Gauteng	Ekurhuleni MM	Densely populated area	Female	57
FD01874598	KwaZulu-Natal	HarryGwala DM	Thinly populated area	Female	0

FD01874599	Western Cape	City of Cape Town MM	Densely populated area	Male	3
FD01874600	Western Cape	City of Cape Town MM	Densely populated area	Female	26
FD01874601	Western Cape	City of Cape Town MM	Densely populated area	Female	34
FD01874603	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Male	2
FD01874604	Gauteng	City of Tshwane MM	Densely populated area	Female	5
FD01874605	Gauteng	City of Tshwane MM	Densely populated area	Female	6
FD01874606	Free State	Mangaung MM	Densely populated area	Male	5
FD01874607	Eastern Cape	Amathole DM	Thinly populated area	Female	35
FD01874608	Western Cape	City of Cape Town MM	Densely populated area	Female	4
FD01874609	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	3
FD01874610	Western Cape	Cape Winelands DM	Intermediate density area	Female	49
FD01874611	Western Cape	City of Cape Town MM	Densely populated area	Female	3
FD01874612	Western Cape	City of Cape Town MM	Densely populated area	Male	33
FD01874614	Eastern Cape	Amathole DM	Thinly populated area	Female	Unknown
FD01874616	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01874617	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	4
FD01874618	Eastern Cape	Amathole DM	Thinly populated area	Female	2
FD01874619	Eastern Cape	Amathole DM	Thinly populated area	Male	5
FD01874620	Eastern Cape	Amathole DM	Thinly populated area	Male	0
FD01874621	Gauteng	Ekurhuleni MM	Densely populated area	Male	3
FD01874623	Western Cape	Garden Route DM	Thinly populated area	Male	1
FD01874624	Western Cape	City of Cape Town MM	Densely populated area	Female	42

FD01874625	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	8
FD01874626	Gauteng	City of Tshwane MM	Densely populated area	Female	80
FD01874627	Western Cape	City of Cape Town MM	Densely populated area	Female	31
FD01874628	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	2
FD01874629	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	4
FD01874630	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	1
FD01874631	Western Cape	City of Cape Town MM	Densely populated area	Female	6
FD01874632	Western Cape	City of Cape Town MM	Densely populated area	Female	25
FD01874633	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	1
FD01874634	Free State	Mangaung MM	Densely populated area	Female	4
FD01874635	Gauteng	West Rand DM	Densely populated area	Female	29
FD01874636	Gauteng	City of Tshwane MM	Densely populated area	Male	3
FD01874637	Gauteng	City of Johannesburg MM	Densely populated area	Female	2
FD01874638	Western Cape	City of Cape Town MM	Densely populated area	Female	9
FD01874640	Gauteng	City of Johannesburg MM	Densely populated area	Male	1
FD01874641	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	Unknown
FD01874642	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Male	13
FD01874643	Free State	Mangaung MM	Densely populated area	Male	Unknown
FD01874644	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	Unknown
FD01874645	KwaZulu-Natal	Ugu DM	Intermediate density area	Male	6
FD01874646	Western Cape	City of Cape Town MM	Densely populated area	Male	3
FD01874647	Gauteng	Ekurhuleni MM	Densely populated area	Female	3

FD01874649	Eastern Cape	Amathole DM	Thinly populated area	Male	Unknown
FD01874650	Western Cape	City of Cape Town MM	Densely populated area	Male	16
FD01874651	Western Cape	West Coast DM	Thinly populated area	Female	74
FD01874652	Eastern Cape	Amathole DM	Thinly populated area	Male	48
FD01874653	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	5
FD01874654	Eastern Cape	Amathole DM	Thinly populated area	Female	63
FD01874655	Gauteng	City of Johannesburg MM	Densely populated area	Female	60
FD01874656	Western Cape	City of Cape Town MM	Densely populated area	Male	11
FD01874657	KwaZulu-Natal	UthukelaDistrict Munincipality	Thinly populated area	Female	74
FD01874658	Gauteng	City of Tshwane MM	Densely populated area	Female	4
FD01874659	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874660	Western Cape	City of Cape Town MM	Densely populated area	Male	18
FD01874661	Western Cape	Cape Winelands DM	Intermediate density area	Male	33
FD01874662	Gauteng	City of Johannesburg MM	Densely populated area	Female	8
FD01874663	Western Cape	City of Cape Town MM	Densely populated area	Male	41
FD01874664	Western Cape	City of Cape Town MM	Densely populated area	Male	15
FD01874666	Gauteng	Ekurhuleni MM	Densely populated area	Male	1
FD01874667	Eastern Cape	Amathole DM	Thinly populated area	Male	3
FD01874668	Free State	Mangaung MM	Densely populated area	Male	4
FD01874669	Limpopo	Waterberg DM	Thinly populated area	Female	38
FD01874670	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01874671	Eastern Cape	Joe Gqabi DM	Thinly populated area	Female	2

FD01874672	Western Cape	City of Cape Town MM	Densely populated area	Male	31
FD01874673	Eastern Cape	Amathole DM	Thinly populated area	Female	2
FD01874675	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	0
FD01874676	Western Cape	West Coast DM	Thinly populated area	Male	2
FD01874677	Western Cape	City of Cape Town MM	Densely populated area	Male	Unknown
FD01874678	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874679	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	3
FD01874680	KwaZulu-Natal	HarryGwala DM	Thinly populated area	Male	4
FD01874681	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874682	Mpumalanga	Ehlanzeni DM	Intermediate density area	UNK	Unknown
FD01874683	Gauteng	City of Johannesburg MM	Densely populated area	Male	7
FD01874684	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	18
FD01874685	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	Unknown
FD01874686	Gauteng	Ekurhuleni MM	Densely populated area	Female	2
FD01874687	Gauteng	City of Johannesburg MM	Densely populated area	Female	53
FD01874688	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	7
FD01874689	North West	Dr Kenneth Kaunda MM	Intermediate density area	Male	36
FD01874691	Mpumalanga	Nkangala DM	Intermediate density area	Male	4
FD01874692	Gauteng	Ekurhuleni MM	Densely populated area	Male	6
FD01874693	Gauteng	City of Tshwane MM	Densely populated area	Female	23
FD01874694	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	2
FD01874696	KwaZulu-Natal	Amajuba DM	Thinly populated area	Female	4

FD01874697	Gauteng	City of Johannesburg MM	Densely populated area	Female	3
FD01874698	KwaZulu-Natal	Ugu DM	Intermediate density area	Male	0
FD01874699	Gauteng	City of Johannesburg MM	Densely populated area	Female	61
FD01874700	Gauteng	Ekurhuleni MM	Densely populated area	Female	30
FD01874701	KwaZulu-Natal	Ugu DM	Intermediate density area	Female	2
FD01874702	Gauteng	City of Johannesburg MM	Densely populated area	Male	4
FD01874703	KwaZulu-Natal	Ugu DM	Intermediate density area	Female	0
FD01874704	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	3
FD01874705	Gauteng	Ekurhuleni MM	Densely populated area	Male	10
FD01874707	Eastern Cape	Sarah Baartman DM	Thinly populated area	Male	7
FD01874708	Free State	Lejweleputswa DM	Intermediate density area	Female	9
FD01874710	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	50
FD01874711	Free State	Mangaung MM	Densely populated area	Female	4
FD01874712	Western Cape	City of Cape Town MM	Densely populated area	Male	7
FD01874713	North West	Dr Kenneth Kaunda MM	Intermediate density area	Male	2
FD01874714	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	8
FD01874715	Free State	Mangaung MM	Densely populated area	Female	3
FD01874716	Gauteng	Ekurhuleni MM	Densely populated area	Male	Unknown
FD01874717	Gauteng	City of Johannesburg MM	Densely populated area	Male	5
FD01874718	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874719	Free State	Mangaung MM	Densely populated area	Female	44
FD01874720	Gauteng	Sedibeng DM	Densely populated area	Male	Unknown

FD01874721	Eastern Cape	Amathole DM	Thinly populated area	Female	7
FD01874722	Gauteng	City of Johannesburg MM	Densely populated area	Male	21
FD01874723	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874724	Gauteng	City of Tshwane MM	Densely populated area	Male	0
FD01874725	Gauteng	City of Tshwane MM	Densely populated area	Male	43
FD01874726	Gauteng	Ekurhuleni MM	Densely populated area	Female	3
FD01874727	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	2
FD01874728	Gauteng	City of Johannesburg MM	Densely populated area	Female	7
FD01874729	Eastern Cape	Amathole DM	Thinly populated area	Male	Unknown
FD01874730	Gauteng	City of Johannesburg MM	Densely populated area	Female	4
FD01874731	Western Cape	Garden Route DM	Thinly populated area	Female	17
FD01874732	Western Cape	City of Cape Town MM	Densely populated area	Female	Unknown
FD01874733	KwaZulu-Natal	Amajuba DM	Thinly populated area	Male	3
FD01874734	Gauteng	City of Tshwane MM	Densely populated area	Male	0
FD01874736	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	1
FD01874737	Eastern Cape	Amathole DM	Thinly populated area	Female	2
FD01874738	Free State	Mangaung MM	Densely populated area	Male	1
FD01874739	Gauteng	City of Tshwane MM	Densely populated area	Male	2
FD01874740	Mpumalanga	Ehlanzeni DM	Intermediate density area	Female	Unknown
FD01874741	Free State	Lejweleputswa DM	Intermediate density area	Male	3
FD01874742	Western Cape	City of Cape Town MM	Densely populated area	Male	28
FD01874743	Eastern Cape	Amathole DM	Thinly populated area	Female	4

FD01874744	Gauteng	Ekurhuleni MM	Densely populated area	Male	32
FD01874745	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874746	Gauteng	West Rand DM	Densely populated area	Male	0
FD01874747	Gauteng	City of Tshwane MM	Densely populated area	Male	1
FD01874748	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	3
FD01874749	Gauteng	City of Johannesburg MM	Densely populated area	Female	9
FD01874750	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Female	1
FD01874751	Gauteng	City of Tshwane MM	Densely populated area	Male	Unknown
FD01874752	Gauteng	City of Johannesburg MM	Densely populated area	Female	2
FD01874753	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01874754	Gauteng	City of Johannesburg MM	Densely populated area	Female	35
FD01874755	Free State	Lejweleputswa DM	Intermediate density area	Male	28
FD01874756	Eastern Cape	Amathole DM	Thinly populated area	Male	0
FD01874757	Gauteng	City of Johannesburg MM	Densely populated area	Female	Unknown
FD01874758	Western Cape	City of Cape Town MM	Densely populated area	Female	32
FD01874760	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	24
FD01874761	Gauteng	City of Tshwane MM	Densely populated area	Female	69
FD01874762	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Female	13
FD01874763	Gauteng	Ekurhuleni MM	Densely populated area	Male	0
FD01874764	KwaZulu-Natal	Amajuba DM	Thinly populated area	Male	3
FD01874766	Gauteng	City of Tshwane MM	Densely populated area	Female	45
FD01874767	Free State	Mangaung MM	Densely populated area	Male	2

FD01874768	Mpumalanga	Gert Sibande DM	Intermediate density area	Female	Unknown
FD01874769	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	29
FD01874770	Gauteng	City of Johannesburg MM	Densely populated area	Female	6
FD01874771	Western Cape	City of Cape Town MM	Densely populated area	Female	55
FD01876599	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	0
FD01876600	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01876601	Northern Cape	Frances Baard DM	Intermediate density area	Female	0
FD01876602	Gauteng	Ekurhuleni MM	Densely populated area	Male	Unknown
FD01876603	Western Cape	City of Cape Town MM	Densely populated area	Male	7
FD01876604	Gauteng	Sedibeng DM	Densely populated area	Female	2
FD01876606	Eastern Cape	Amathole DM	Thinly populated area	Female	5
FD01876607	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01876609	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Female	74
FD01876610	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	0
FD01876611	Western Cape	City of Cape Town MM	Densely populated area	Male	7
FD01876612	Eastern Cape	Amathole DM	Thinly populated area	Male	1
FD01876613	Western Cape	City of Cape Town MM	Densely populated area	Female	86
FD01876615	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01876617	Eastern Cape	Unknown	Unknown	Male	2
FD01876618	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	26
FD01876619	Western Cape	City of Cape Town MM	Densely populated area	Male	11
FD01876620	Gauteng	Unknown	Unknown	Female	Unknown
FD01876621	Western Cape	City of Cape Town MM	Densely populated area	Female	22

FD01876622	Eastern Cape	Amathole DM	Thinly populated area	Female	46
FD01876623	Eastern Cape	Unknown	Unknown	Male	0
FD01876624	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01876626	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01876627	Western Cape	City of Cape Town MM	Densely populated area	UNKNOWN	1
FD01876628	Western Cape	City of Cape Town MM	Densely populated area	Male	21
FD01876629	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	5
FD01876630	Western Cape	City of Cape Town MM	Densely populated area	Female	5
FD01876631	Western Cape	City of Cape Town MM	Densely populated area	Male	41
FD01876632	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01876633	Mpumalanga	Gert Sibande DM	Intermediate density area	Female	38
FD01876635	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Female	4
FD01876636	Gauteng	Ekurhuleni MM	Densely populated area	Female	Unknown
FD01876637	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	2
FD01876638	Western Cape	City of Cape Town MM	Densely populated area	Male	1
FD01876639	Western Cape	Garden Route DM	Thinly populated area	Male	3
FD01876640	Western Cape	West Coast DM	Thinly populated area	Male	76
FD01876641	Western Cape	City of Cape Town MM	Densely populated area	Male	9
FD01876642	Western Cape	City of Cape Town MM	Densely populated area	Female	4
FD01876643	Eastern Cape	Amathole DM	Thinly populated area	Female	3
FD01876644	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01876645	Western Cape	City of Cape Town MM	Densely populated area	Female	1
FD01876647	Northern Cape	Frances Baard DM	Intermediate density area	Male	Unknown

FD01876648	Eastern Cape	Unknown	Unknown	Male	45
FD01876649	Western Cape	City of Cape Town MM	Densely populated area	Male	0
FD01876650	Gauteng	City of Tshwane MM	Densely populated area	Female	3
FD01876651	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	13
FD01876652	Western Cape	City of Cape Town MM	Densely populated area	Female	43
FD01876653	Western Cape	City of Cape Town MM	Densely populated area	Male	4
FD01876654	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Female	30
FD01876655	Gauteng	Sedibeng DM	Densely populated area	Male	2
FD01876656	KwaZulu-Natal	Unknown	Unknown	Female	4
FD01876658	Gauteng	Sedibeng DM	Densely populated area	Male	1
FD01876660	KwaZulu-Natal	uMgungundlovu DM	Intermediate density area	Male	8
FD01876661	Gauteng	City of Tshwane MM	Densely populated area	Female	4
FD01876662	Western Cape	Overberg DM	Thinly populated area	Male	1
FD01876663	Eastern Cape	Amathole DM	Thinly populated area	Male	Unknown
FD01876664	Western Cape	West Coast DM	Thinly populated area	Male	1
FD01876665	Free State	Mangaung MM	Densely populated area	Male	2
FD01876666	Western Cape	City of Cape Town MM	Densely populated area	Female	2
FD01876668	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	Unknown
FD01876669	Eastern Cape	Nelson Mandela Bay MM	Densely populated area	Male	30
FD01876670	KwaZulu-Natal	Uthukela District Munincipality	Thinly populated area	Male	1
FD01876671	KwaZulu-Natal	Ugu DM	Intermediate density area	Female	3
FD01876672	Gauteng	Unknown	Unknown	Male	Unknown
FD01876673	Western Cape	City of Cape Town MM	Densely populated area	Female	2

FD01876674	Western Cape	City of Cape Town MM	Densely populated area	Male	2
FD01876675	Western Cape	Unknown	Unknown	Female	3
FD01876676	Eastern Cape	Amathole DM	Thinly populated area	Female	3
FD01876677	Gauteng	City of Tshwane MM	Densely populated area	Female	2
FD01876678	Gauteng	Unknown	Unknown	Female	Unknown
FD01876679	Gauteng	City of Johannesburg MM	Densely populated area	Male	10
FD01876680	Gauteng	City of Johannesburg MM	Densely populated area	Female	24
FD01876681	Gauteng	Unknown	Unknown	Male	0
FD01876682	Eastern Cape	Chris Hani DM	Thinly populated area	Female	3
FD01876683	Western Cape	Unknown	Unknown	Female	24
FD01876684	Eastern Cape	Amathole DM	Thinly populated area	Male	Unknown
FD01876685	Gauteng	City of Johannesburg MM	Densely populated area	Male	Unknown
FD01876686	KwaZulu-Natal	City of eThekwini MM	Densely populated area	Male	5
FD01876687	Western Cape	West Coast DM	Thinly populated area	Female	24
FD01876688	Western Cape	City of Cape Town MM	Densely populated area	Male	0
FD01876689	Free State	Mangaung MM	Densely populated area	Female	2
FD01876690	Mpumalanga	Gert Sibande DM	Intermediate density area	Female	0
FD01876691	Western Cape	West Coast DM	Thinly populated area	Female	1
FD01876692	Western Cape	Unknown	Unknown	Male	37
FD01876695	Western Cape	Unknown	Unknown	Female	7

Table 5. Phenotypic resistance to a selection of antimicrobials and the amino acids at two positions in two genes in the quinolone resistance determining region, extracted in silico from the assembled genomes.

Isolate	Ampicillin	Chloramphenicol	Streptomycin	Tetracycline	Nalidixic acid	Ciprofloxacin	Ceftriaxone	Cotrimoxazole	*gyrA* p83 & p89	*parC* p80 & p91
FD01872847	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872848	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872849	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872850	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872851	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872852	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872853	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872854	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872855	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872856	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872857	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872858	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872859	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872860	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872861	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872863	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872865	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872866	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872867	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872868	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872869	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872870	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872871	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872872	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872873	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872874	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872875	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ

FD01872876	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872877	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872878	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872879	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872880	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872881	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872882	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872883	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872884	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872885	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872887	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872888	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872889	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872891	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872892	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872893	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872894	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01872895	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872896	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872897	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872898	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872899	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872900	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872901	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872902	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01872903	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872904	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872905	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872906	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872907	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872908	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01872909	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872910	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01872911	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872912	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872913	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872914	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872915	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872916	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872917	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872918	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01872919	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872920	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872921	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872922	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872924	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872925	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872926	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01872927	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872928	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872929	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872930	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872931	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872932	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872933	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872934	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01872935	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872936	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872938	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872939	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01872940	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01872941	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01872942	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01873906	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873907	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873908	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873909	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873911	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873914	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873915	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873916	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873917	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873918	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873919	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873920	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873922	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01873923	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873925	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873926	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873927	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873928	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873930	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873931	Resistant	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873932	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873933	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873934	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01873935	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873936	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873938	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873939	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01873940	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873941	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873942	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01873943	Resistant	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873944	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873946	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873947	Susceptible	Susceptible	Resistant	Resistant	Resistant	Susceptible	Susceptible	Resistant	SY	SQ
FD01873948	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873949	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873950	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873951	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873952	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873954	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873955	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01873956	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873957	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873958	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SY	SQ
FD01873959	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873962	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	LD	SQ
FD01873963	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873964	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873965	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873966	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873967	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873968	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873970	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873971	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873972	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873973	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Resistant	Resistant	SD	SQ
FD01873974	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873975	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873976	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873978	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873979	Susceptible	Susceptible	Resistant	Resistant	Resistant	Susceptible	Susceptible	Resistant	LD	SQ

FD01873980	Resistant	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873982	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873983	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873986	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873987	Resistant	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873988	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873990	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873991	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873994	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873995	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873996	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873998	Susceptible	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01873999	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874098	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874099	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874100	Susceptible	Untested	Susceptible	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874101	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874102	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874103	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874104	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874105	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874106	Susceptible	Intermediate	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874107	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874108	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874109	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874110	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874111	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874113	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874114	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Resistant	Resistant	SD	SQ
FD01874115	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874116	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ

FD01874117	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874118	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874119	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874120	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874121	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874123	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874124	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874125	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874126	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874127	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874128	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874129	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874130	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874131	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874132	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874133	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874134	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874135	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874136	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874137	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874138	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874139	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874140	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874141	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874142	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874143	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874144	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874145	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874146	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874147	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874148	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ

FD01874149	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874150	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874151	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874152	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874155	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874156	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874157	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874158	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874159	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874160	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874161	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874162	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874163	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874164	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874165	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874166	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874167	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874168	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874169	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874170	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874171	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874172	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874174	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874175	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874176	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874179	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874180	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874181	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874182	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874183	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874184	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01874185	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874186	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874187	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874189	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874190	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874191	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874192	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874193	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874579	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874580	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874581	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874582	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874583	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874585	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874587	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874588	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874589	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874590	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874591	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874592	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874593	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874594	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874595	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874596	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874597	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874598	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874599	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874600	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874601	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874603	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874604	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ

FD01874605	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874606	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874607	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874608	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874609	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874610	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874611	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874612	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874614	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874616	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874617	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874618	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874619	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874620	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874621	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874623	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874624	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874625	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874626	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874627	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874628	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874629	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874630	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874631	Susceptible	Untested	Untested	Untested	Untested	Resistant	Untested	Untested	SD	SQ
FD01874632	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874633	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874634	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874635	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874636	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874637	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874638	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ

FD01874640	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874641	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874642	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874643	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874644	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874645	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874646	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874647	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874649	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874650	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874651	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874652	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874653	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874654	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874655	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874656	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874657	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874658	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874659	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874660	Untested	Untested	Untested	Untested	Untested	Untested	Untested	Untested	SD	SQ
FD01874661	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874662	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874663	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874664	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874666	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874667	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874668	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874669	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874670	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874671	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874672	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ

FD01874673	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874675	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874676	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874677	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874678	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874679	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874680	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874681	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Untested	Susceptible	Resistant	SD	SQ
FD01874682	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874683	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874684	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874685	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874686	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874687	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874688	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874689	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874691	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874692	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874693	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874694	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874696	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874697	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874698	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874699	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874700	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874701	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874702	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874703	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874704	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874705	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874707	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01874708	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874710	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	LG	IQ
FD01874711	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874712	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874713	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874714	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874715	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874716	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874717	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874718	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874719	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874720	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874721	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874722	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874723	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874724	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874725	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874726	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874727	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SG	SQ
FD01874728	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874729	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874730	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874731	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874732	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874733	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874734	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874736	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874737	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874738	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874739	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874740	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ

FD01874741	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874742	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874743	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874744	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874745	Susceptible	Susceptible	Resistant	Resistant	Resistant	Susceptible	Susceptible	Resistant	SG	SQ
FD01874746	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874747	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874748	Resistant	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874749	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874750	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874751	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874752	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874753	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874754	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874755	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874756	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874757	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874758	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874760	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874761	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01874762	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874763	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874764	Susceptible	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874766	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874767	Untested	Untested	Untested	Untested	Untested	Susceptible	Untested	Untested	SD	SQ
FD01874768	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874769	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874770	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01874771	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876599	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876600	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01876601	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876602	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876603	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876604	Resistant	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876606	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876607	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876609	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876610	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876611	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876612	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876613	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876615	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876617	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876618	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876619	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876620	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876621	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876622	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876623	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876624	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876626	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876627	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876628	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876629	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876630	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876631	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876632	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876633	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876635	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876636	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876637	Susceptible	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01876638	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876639	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876640	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876641	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876642	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876643	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876644	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876645	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876647	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876648	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876649	Susceptible	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876650	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876651	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876652	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876653	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876654	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876655	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876656	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876658	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876660	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876661	Resistant	Susceptible	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876662	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876663	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876664	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876665	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876666	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876668	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876669	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876670	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876671	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876672	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

FD01876673	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876674	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876675	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876676	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876677	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876678	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876679	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876680	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876681	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876682	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876683	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876684	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876685	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876686	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876687	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876688	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876689	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876690	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	SD	SQ
FD01876691	Resistant	Susceptible	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876692	Resistant	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ
FD01876695	Resistant	Resistant	Resistant	Susceptible	Susceptible	Susceptible	Susceptible	Resistant	SD	SQ

Table 6. South African isolate BAPS cluster (Chapters 4 and 5) and SonneiTyping prediction for S. sonnei isolates (right) (Chapter 5).

Isolate	BAPS cluster	Isolate	SonneiTyping prediction	BAPS cluster
FD01872847	1	FD01873906	3.7.11	1
FD01872848	1	FD01873907	3.7.7	2
FD01872849	1	FD01873908	3.7.11	1
FD01872850	2	FD01873909	3.7.9	3
FD01872851	2	FD01873911	3.7.11	1
FD01872852	3	FD01873914	3.7.11	1
FD01872853	2	FD01873915	3.7.9	3
FD01872854	2	FD01873916	3.7.11	1
FD01872855	2	FD01873917	3.7.11	1
FD01872856	2	FD01873918	3.7.9	3
FD01872857	1	FD01873919	3.7.11	1
FD01872858	1	FD01873920	3.7.11	1
FD01872859	1	FD01873922	3.7.9	3
FD01872860	3	FD01873923	3.7.11	1
FD01872861	2	FD01873925	3.7.11	1
FD01872863	1	FD01873926	3.7.11	1
FD01872865	2	FD01873927	3.7.9	3
FD01872866	2	FD01873928	3.7.9	3
FD01872867	3	FD01873930	3.7.11	1
FD01872868	3	FD01873931	3.7.7	2
FD01872869	4	FD01873932	3.7.11	1
FD01872870	3	FD01873933	3.7.9	3
FD01872871	2	FD01873934	3.7.11	1
FD01872872	3	FD01873935	3.7.11	1
FD01872873	2	FD01873936	3.7.9	3
FD01872874	2	FD01873938	2.8.2	1
FD01872875	2	FD01873939	3.7.9	3
FD01872876	1	FD01873940	3.7.11	1
FD01872877	3	FD01873941	3.7.11	1
FD01872878	2	FD01873942	3.7.11	1
FD01872879	4	FD01873943	3.7.7	2
FD01872880	4	FD01873944	-	1
FD01872881	3	FD01873946	3.7.11	1
FD01872882	4	FD01873947	3.7.9	3
FD01872883	1	FD01873948	3.7.9	3
FD01872884	1	FD01873949	3.7.9	3
FD01872885	1	FD01873950	3.7.9	3
FD01872887	2	FD01873951	3.7.11	1
FD01872888	2	FD01873952	3.7.11	1
FD01872889	4	FD01873954	3.7.11	1
FD01872891	1	FD01873955	3.7.9	3

FD01872892	2		FD01873956	3.7.11	1
FD01872893	3		FD01873957	3.7.11	1
FD01872894	1		FD01873958	3.7.9	3
FD01872895	2		FD01873959	3.7.11	1
FD01872896			FD01873962	3.6.1	2
FD01872897	1		FD01873963	3.7.7	2
FD01872898	3		FD01873964	3.7.9	3
FD01872899	3		FD01873965	3.7.9	3
FD01872900	4		FD01873966	3.7.11	1
FD01872901	2		FD01873967	3.7.11	1
FD01872902	3		FD01873968	3.7.11	1
FD01872903	2		FD01873970	3.7.11	1
FD01872904	2		FD01873971	3.7.11	1
FD01872905	4		FD01873972	3.7.11	1
FD01872906	4		FD01873973	3.7.9	3
FD01872907	1		FD01873974	3.7.9	3
FD01872908	2		FD01873975	3.7.9	3
FD01872909	4		FD01873976	3.7.11	1
FD01872910	2		FD01873978	3.7.11	1
FD01872911	1		FD01873979	3.7.9	3
FD01872912	3		FD01873980	3.7.7	2
FD01872913	1		FD01873982	2.8.2	2
FD01872914	2		FD01873983	3.7.9	3
FD01872915	2		FD01873986	3.7.11	1
FD01872916	3		FD01873987	3.7.7	2
FD01872917	2		FD01873988	3.7.7	2
FD01872918	2		FD01873990	3.7.11	1
FD01872919	1		FD01873991	3.7.9	3
FD01872920	2		FD01873994	3.7.11	1
FD01872921	4		FD01873995	3.7.11	1
FD01872922	1		FD01873996	3.7.11	1
FD01872924	4		FD01873998	3.7.11	1
FD01872925	1		FD01873999	3.7.9	3
FD01872926	3		FD01874098	3.7.9	3
FD01872927	3		FD01874099	3.7.9	3
FD01872928	2		FD01874100	3.7.7	2
FD01872929	4		FD01874101	3.7.11	1
FD01872930	3		FD01874102	3.7.11	1
FD01872931	3		FD01874103	3.7.9	3
FD01872932	1		FD01874104	3.7.7	2
FD01872933	2		FD01874105	2.8.2	4
FD01872934	4		FD01874106	3.7.11	1
FD01872935	2		FD01874107	3.7.11	1
FD01872936	1		FD01874108	3.7.11	1
FD01872938	3		FD01874109	3.7.9	3

FD01872939	2		FD01874110	3.7.9	3
FD01872940	3		FD01874111	3.7.9	3
FD01872941	1		FD01874113	3.7.11	1
FD01872942	3		FD01874114	3.7.11	1
FD01874580	3		FD01874115	2.8.2	4
FD01874581	3		FD01874116	3.7.9	3
FD01874582	1		FD01874117	3.7.9	3
FD01874583	2		FD01874118	3.7.11	1
FD01874585	3		FD01874119	3.7.11	1
FD01874587	1		FD01874120	2.8.2	4
FD01874588	4		FD01874121	3.7.11	1
FD01874589	1		FD01874123	3.7.9	3
FD01874590	2		FD01874124	3.7.11	1
FD01874591	4		FD01874125	3.7.11	1
FD01874592	2		FD01874126	3.7.9	3
FD01874593	2		FD01874127	3.7.11	1
FD01874594	4		FD01874128	3.7.9	3
FD01874595	2		FD01874129	3.7.7	2
FD01874596	3		FD01874130	3.7.9	3
FD01874597	2		FD01874131	3.7.11	1
FD01874598	2		FD01874132	3.7.11	1
FD01874599	4		FD01874133	3.7.9	3
FD01874600	3		FD01874134	3.7.9	3
FD01874601	1		FD01874135	3.7.9	3
FD01874603	2		FD01874136	3.7.11	1
FD01874604	1		FD01874137	3.7.11	1
FD01874605	2		FD01874138	3.7.9	3
FD01874606	1		FD01874139	3.7.9	3
FD01874607	2		FD01874140	3.7.9	3
FD01874608	1		FD01874141	3.7.11	1
FD01874609	4		FD01874142	3.7.9	3
FD01874610	4		FD01874143	3.7.11	1
FD01874611	2		FD01874144	3.7.11	1
FD01874612	1		FD01874145	3.7.11	1
FD01874614	2		FD01874146	3.7.7	2
FD01874616	2		FD01874147	3.7.9	3
FD01874617	3		FD01874148	2.8.2	2
FD01874618	4		FD01874149	3.7.9	3
FD01874619	2		FD01874150	3.7.11	1
FD01874620	2		FD01874151	3.7.9	3
FD01874621	4		FD01874152	3.7.11	1
FD01874623	4		FD01874155	3.7.9	3
FD01874624	2		FD01874156	3.7.11	1
FD01874625	3		FD01874157	3.7.11	1
FD01874626	1		FD01874158	3.7.9	3

FD01874627	4		FD01874159	3.6.0	2
FD01874628	2		FD01874160	3.7.11	1
FD01874629	4		FD01874161	3.7.11	1
FD01874630	2		FD01874162	3.7.9	3
FD01874631	3		FD01874163	3.7.11	1
FD01874632	3		FD01874164	3.7.9	3
FD01874633	3		FD01874165	3.7.9	3
FD01874634	1		FD01874166	3.7.9	3
FD01874635	1		FD01874167	3.7.11	1
FD01874636	3		FD01874168	3.7.11	1
FD01874637	4		FD01874169	3.7.7	2
FD01874638	3		FD01874170	3.7.9	3
FD01874640	2		FD01874171	3.7.9	3
FD01874641	2		FD01874172	3.7.11	1
FD01874642	2		FD01874174	3.7.11	1
FD01874643	3		FD01874175	3.7.9	3
FD01874644	3		FD01874176	3.7.11	1
FD01874645	1		FD01874179	3.7.11	1
FD01874646	2		FD01874180	3.7.9	3
FD01874647	3		FD01874181	3.7.9	3
FD01874649	4		FD01874182	2.8.2	2
FD01874650	4		FD01874183	3.7.11	1
FD01874651	4		FD01874184	3.7.11	1
FD01874652	2		FD01874185	3.7.11	1
FD01874653	2		FD01874186	3.7.9	3
FD01874654	4		FD01874187	3.7.11	1
FD01874655	1		FD01874189	3.7.9	3
FD01874656	4		FD01874190	3.7.7	2
FD01874657	2		FD01874191	3.7.11	1
FD01874658	1		FD01874192	3.7.7	2
FD01874659	2		FD01874193	3.7.7	2
FD01874660	4		FD01874579		2
FD01874661	4		FD01874667	3.7.9	3
FD01874662	2		FD01874675	3.7.9	3
FD01874663	2		FD01874676	3.7.9	3
FD01874664	4		FD01874677	3.7.11	1
FD01874666	1		FD01874678	3.7.7	2
FD01874668	1		FD01874679	3.7.9	3
FD01874669	1		FD01874680	3.7.9	3
FD01874670	2		FD01874681	3.7.11	1
FD01874671	4		FD01874682	3.7.11	1
FD01874672	4		FD01874683	3.7.7	2
FD01874673	4		FD01874684	3.7.9	3
FD01876599	2		FD01874685	3.7.11	1
FD01876600	1		FD01874686	3.7.11	1

FD01876601	1		FD01874687	3.4.1	
FD01876602	2		FD01874688	3.7.11	1
FD01876603	1		FD01874689	3.7.11	1
FD01876604	1		FD01874691	3.7.7	2
FD01876606	1		FD01874692	3.7.11	1
FD01876607	2		FD01874693	3.7.11	1
FD01876609	1		FD01874694	3.7.9	3
FD01876610	2		FD01874696	3.7.11	1
FD01876611	2		FD01874697	3.7.11	1
FD01876612	4		FD01874698	3.7.9	3
FD01876613	2		FD01874699	3.7.11	1
FD01876615	2		FD01874700	3.7.11	1
FD01876617	4		FD01874701	3.7.11	1
FD01876618	2		FD01874702	3.7.9	3
FD01876619	1		FD01874703	3.7.11	1
FD01876620	1		FD01874704	3.7.9	3
FD01876621	2		FD01874705	3.7.11	1
FD01876622	3		FD01874707	3.7.9	3
FD01876623	4		FD01874708	3.7.9	3
FD01876624	2		FD01874710	3.6.1.1.3	2
FD01876626	1		FD01874711	3.7.11	1
FD01876627	4		FD01874712	3.7.9	3
FD01876628	2		FD01874713	3.7.9	3
FD01876629	4		FD01874714	3.7.9	3
FD01876630	1		FD01874715	3.7.11	1
FD01876631	4		FD01874716	3.7.11	1
FD01876632	1		FD01874717	3.7.11	1
FD01876633	2		FD01874718	3.7.11	1
FD01876635	1		FD01874719	3.7.9	3
FD01876636	3		FD01874720	3.7.9	3
FD01876637	2		FD01874721	3.7.7	2
FD01876638	4		FD01874722	3.7.9	3
FD01876639	1		FD01874723	2.8.2	2
FD01876640	2		FD01874724	3.7.9	3
FD01876641	3		FD01874725	3.7.9	3
FD01876642	1		FD01874726	3.7.9	3
FD01876643	4		FD01874727	3.7.11	1
FD01876644	1		FD01874728	3.7.11	1
FD01876645	3		FD01874729	3.7.11	1
FD01876647	4		FD01874730	3.7.9	3
FD01876648	3		FD01874731	2.8.2	
FD01876649	2		FD01874732	3.7.9	3
FD01876650	1		FD01874733	3.7.9	3
FD01876651	3		FD01874734	3.7.11	1
FD01876652	3		FD01874736	3.7.9	3

FD01876653	2
FD01876654	2
FD01876655	3
FD01876656	3
FD01876658	1
FD01876660	1
FD01876661	1
FD01876662	4
FD01876663	2
FD01876664	2
FD01876665	1
FD01876666	4
FD01876668	3
FD01876669	4
FD01876670	2
FD01876671	1
FD01876672	
FD01876673	3
FD01876674	1
FD01876675	2
FD01876676	4
FD01876677	3
FD01876678	4
FD01876679	2
FD01876680	2
FD01876681	3
FD01876682	2
FD01876683	1
FD01876684	2
FD01876685	2
FD01876686	2
FD01876687	3
FD01876688	4
FD01876689	1
FD01876690	2
FD01876691	3
FD01876692	4
FD01876695	3

FD01874737	3.7.7	2
FD01874738	3.7.9	3
FD01874739	3.7.9	3
FD01874740	3.7.9	3
FD01874741	3.7.11	1
FD01874742	3.7.11	1
FD01874743	3.7.9	3
FD01874744	3.7.11	1
FD01874745	3.7.11	1
FD01874746	3.7.9	3
FD01874747	3.7.11	1
FD01874748	3.7.11	1
FD01874749	3.7.9	3
FD01874750	3.7.9	3
FD01874751	3.7.11	1
FD01874752	3.7.11	1
FD01874753	3.7.11	1
FD01874754	3.7.11	1
FD01874755	3.7.11	1
FD01874756	3.7.11	1
FD01874757	3.7.9	3
FD01874758	3.7.11	1
FD01874760	3.7.9	3
FD01874761	3.7.11	1
FD01874762	3.7.9	3
FD01874763	3.7.9	3
FD01874764	3.7.11	1
FD01874766	2.8.2	2
FD01874767	3.7.7	2
FD01874768	3.7.9	3
FD01874769	3.7.11	1
FD01874770	3.7.9	3
FD01874771	3.7.11	1

Table 7. GEMS study isolate accession numbers, reported and predicted serotypes, phylotype prediction (S. sonnei), sample date and country and patient age (Chapter 6).

Isolate	Accession	Reported serotype	SonneiTyping prediction	Shigatyper serotype prediction	Re-sequenced?	Sample date	Country	Age (months)
FD01843750	ERR6005285	*S. flexneri* 2b	3.7.17	S. sonnei form II	No	08/06/2010	The Gambia	9
FD01843766	ERR6005681	*S. sonnei*	3.7.17	S. sonnei form II	No	17/06/2010	The Gambia	22
FD01843806	ERR6005682	*S. sonnei*	3.7.16	S. sonnei form II	No	06/06/2010	The Gambia	10
FD01844496	ERR6005612	*S. sonnei*	2.8.2	S. sonnei form II	No	27/07/2009	Kenya	15
FD01844520	ERR6005613, ERR6006233	*S. sonnei*	2.8.2	S. sonnei form II	Yes	14/10/2009	Kenya	10
FD01844527	ERR6005610	*S. sonnei*	3.7.17	S. sonnei form II	No	31/12/2009	The Gambia	9
FD01844530	ERR6005619	*S. sonnei*	3.7.12	S. sonnei form II	No	22/02/2010	Kenya	15
FD01844545	ERR6005617	*S. sonnei*	3.7.10	S. sonnei form II	No	11/02/2010	Kenya	10
FD01844553	ERR6005618	*S. sonnei*	2.8.2	S. sonnei form II	No	16/02/2010	Kenya	44
FD01844557	ERR6005606	*S. sonnei*	3.7.17	S. sonnei form II	No	18/09/2009	The Gambia	22
FD01844558	ERR6005608	*S. sonnei*	3.7.17	S. sonnei form II	No	10/09/2009	The Gambia	27
FD01844565	ERR6005607	*S. sonnei*	3.7.17	S. sonnei form II	No	04/10/2009	The Gambia	28
FD01845174	ERR6005624	*S. sonnei*	3.7.7	S. sonnei form II	No	29/01/2010	Mozambique	17
FD01845182	ERR6005625	*S. sonnei*	2.8.2	S. sonnei form II	No	24/12/2009	Mozambique	33
FD01845740	ERR6005673	*S. sonnei*	3.7.16	S. sonnei form II	No	23/08/2010	Mali	23
FD01845749	ERR6005267	*S. flexneri* 3a	3.7.10	S. sonnei form II	No	05/07/2010	Kenya	27
FD01845753	ERR6005663	*S. sonnei*	2.8.2	S. sonnei form II	No	08/07/2008	Mali	22
FD01845755	ERR6005667	*S. sonnei*	3.7.17	S. sonnei form II	No	22/04/2010	Mali	15
FD01845763	ERR6005668	*S. sonnei*	3.7.16	S. sonnei form II	No	03/05/2010	Mali	26
FD01845771	ERR6005669	*S. sonnei*	3.7.16	S. sonnei (form I)	No	02/06/2010	Mali	32
FD01845795	ERR6005672	*S. sonnei*	3.7.16	S. sonnei form II	No	21/04/2009	Mali	31
FD01845802	ERR6005665	*S. sonnei*	3.7.16	S. sonnei form II	No	22/04/2009	Mali	14
FD01845810	ERR6005666	*S. sonnei*	3.7.16	S. sonnei form II	No	03/08/2009	Mali	22
FD01845812	ERR6005676, ERR6006237	*S. sonnei*	2.8.2	S. sonnei form II	Yes	19/05/2010	Kenya	15

FD01845813	ERR6005679	*S. sonnei*	3.7.8	S. sonnei form II	No	17/12/2008	Kenya	35
FD01847238	ERR6005602, ERR6006190	*S. sonnei*	3.7.17	S. sonnei form II	Yes	21/07/2009	The Gambia	43
FD01847261	ERR6005599	*S. sonnei*	3.7.17	S. sonnei form II	No	13/04/2009	The Gambia	19
FD01874204	ERR6005451	*S. sonnei*	3.7.10	S. sonnei form II	No	18/02/2008	Kenya	20
FD01874206	ERR6005459	*S. sonnei*	2.8.2	S. sonnei form II	No	06/05/2008	Kenya	43
FD01874219	ERR6005446	*S. sonnei*	3.7.17	S. sonnei form II	No	15/07/2008	The Gambia	19
FD01874243	ERR6005447	*S. sonnei*	3.7.17	S. sonnei form II	No	07/08/2008	The Gambia	8
FD01874244	ERR6005449	*S. sonnei*	2.8.2	S. sonnei form II	No	24/06/2008	Mali	24
FD01874247	ERR6005461	*S. sonnei*	3.7.10	S. sonnei form II	No	03/07/2008	Kenya	5
FD01874251	ERR6005448	*S. sonnei*	3.7.17	S. sonnei (form I)	No	09/08/2008	The Gambia	21
FD01874261	ERR6005452	*S. sonnei*	3.7.10	S. sonnei form II	No	12/03/2008	Kenya	44
FD01874333	ERR6005512	*S. sonnei*	3.7.17	S. sonnei form II	No	06/01/2009	The Gambia	16
FD01876350	ERR6005440	*S. sonnei*	3.7.17	S. sonnei form II	No	07/05/2008	The Gambia	17
FD01876358	ERR6005441	*S. sonnei*	3.7.17	S. sonnei form II	No	12/05/2008	The Gambia	22
FD01876364	ERR6005431	*S. sonnei*	2.1.1,3.7.17	S. sonnei (form I)	No	15/12/2007	The Gambia	14
FD01876366	ERR6005442	*S. sonnei*	3.6.1	S. sonnei form II	No	14/07/2008	The Gambia	23
FD01876373	ERR6005436	*S. sonnei*	3.7.17	S. sonnei form II	No	25/03/2008	The Gambia	52
FD01876390	ERR6005443	*S. sonnei*	3.7.17	S. sonnei (form I)	No	03/07/2008	The Gambia	19
FD01876396	ERR6005432	*S. sonnei*	3.7.17,2.2.1	S. sonnei form II	No	02/02/2008	The Gambia	15
FD01876397	ERR6005437	*S. sonnei*	3.7.17	S. sonnei form II	No	28/03/2008	The Gambia	29
FD01876544	ERR6005532	*S. sonnei*	3.7.10	S. sonnei (form I)	No	25/02/2009	Kenya	36
FD01876552	ERR6005533	*S. sonnei*	2.8.2	S. sonnei form II	No	27/03/2009	Kenya	21
FD01876569	ERR6005538	*S. sonnei*	3.7.7	S. sonnei form II	No	01/04/2008	Mozambique	27
FD01876576	ERR6005536	*S. sonnei*	3.7.10	S. sonnei form II	No	19/05/2009	Kenya	20
FD01876577	ERR6005539	*S. sonnei*	2.8.2	S. sonnei form II	No	19/03/2008	Mozambique	29
FD01843719	ERR6005290	*S. flexneri 6*		S. flexneri 6	No	07/06/2010	The Gambia	15
FD01843726	ERR6005283	*S. flexneri 1b*		S. flexneri 1b	No	27/05/2010	The Gambia	18
FD01843730	ERR6005313	*S. flexneri 2a*		S. flexneri 2a	No	27/09/2010	Mozambique	21

FD01843735	ERR6005292	*S. flexneri 2a*	S. flexneri 2a	No	03/07/2010	The Gambia	28
FD01843741	ERR6005275	*S. flexneri 2a*	S. flexneri 2a	No	18/04/2010	The Gambia	20
FD01843742	ERR6005284	*S. flexneri 2a*	S. flexneri 2a	No	08/06/2010	The Gambia	12
FD01843749	ERR6005276	*S. flexneri 2b*	S. flexneri 2b	No	22/04/2010	The Gambia	12
FD01843751	ERR6005293	*S. flexneri Y*	S. flexneri Y	No	06/08/2010	The Gambia	18
FD01843758	ERR6005680	*S. sonnei*	S. flexneri 2b	No	16/06/2010	The Gambia	19
FD01843759	ERR6005294	*S. flexneri 1b*	S. flexneri 1b	No	13/07/2010	The Gambia	18
FD01843773	ERR6005278	*S. flexneri X*	S. flexneri Xv (4c)	No	06/05/2010	The Gambia	15
FD01843782	ERR6005287	*S. flexneri 2a*	S. flexneri 2a	No	18/06/2010	The Gambia	7
FD01843790	ERR6005288	*S. flexneri 2a*	S. flexneri 2a	No	06/08/2010	The Gambia	18
FD01843797	ERR6005281	*S. flexneri 3a*	S. flexneri 3a	No	24/05/2010	The Gambia	11
FD01843798	ERR6005289	*S. flexneri 2a*	S. flexneri 2a	No	24/06/2010	The Gambia	25
FD01843801	ERR6005311	*S. flexneri 1b*	S. flexneri 1b	No	05/11/2010	Mozambique	21
FD01843805	ERR6005282	*S. flexneri 2a*	S. flexneri 2a	No	24/04/2010	The Gambia	12
FD01844295	ERR6005392, ERR6006215	*S. flexneri 1b*	S. flexneri 1b	Yes	26/10/2009	Mozambique	25
FD01844303	ERR6005393, ERR6006214	*S. flexneri X*	S. flexneri Xv (4c)	Yes	18/09/2009	Mozambique	29
FD01844335	ERR6005396	*S. flexneri 7a*	S. flexneri,	No	23/10/2008	Kenya	5
FD01844351	ERR6005397	*S. flexneri 3a*	S. flexneri,	No	24/05/2010	Kenya	10
FD01844357	ERR6005383, ERR6006247	*S. flexneri 3a*	S. flexneri 3a	Yes	16/09/2010	Kenya	2
FD01844382	ERR6005391	*S. flexneri 2a*	S. flexneri 2av	No	10/08/2009	Mozambique	24
FD01844488	ERR6005116	*S. flexneri 1b*	S. flexneri 1b	No	15/07/2009	Kenya	5
FD01844489	ERR6005125	*S. flexneri 1b*	S. flexneri 1b	No	25/01/2010	Kenya	33
FD01844493	ERR6005093	*S. flexneri 6*	S. flexneri 6	No	16/07/2009	The Gambia	25
FD01844494	ERR6005100, ERR6006192	*S. flexneri 1b*	S. flexneri 1b	Yes	10/09/2009	The Gambia	28
FD01844495	ERR6005110	*S. flexneri 3a*	S. flexneri 3a	No	27/12/2009	The Gambia	26

FD01844497	ERR6005126	*S. flexneri 3a*	S. flexneri 3a	No	28/01/2010	Kenya	16
FD01844501	ERR6005094	*S. flexneri 2a*	S. flexneri 2a	No	07/07/2009	The Gambia	13
FD01844502	ERR6005101	*S. flexneri 1b*	S. flexneri 1b	No	08/10/2009	The Gambia	45
FD01844503	ERR6005111	*S. flexneri 2a*	S. flexneri 2a	No	27/12/2009	The Gambia	23
FD01844504	ERR6005117	*S. flexneri 3a*	S. flexneri 3a	No	09/09/2009	Kenya	47
FD01844505	ERR6005127	*S. flexneri 3a*	S. flexneri 3a	No	02/02/2010	Kenya	24
FD01844506	ERR6005133	*S. flexneri 6*	S. flexneri 6	No	12/04/2010	Kenya	5

Supplementary Figures

Removed from publicly available version

Figure 15. The thirty-five page Standard Operating Procedures document for the biochemical identification and antimicrobial resistance testing of bacterial isolates collected as part of public healthcare surveillance in South Africa by the Group for Enteric, Respiratory and Meningeal Diseases Surveillance in South Africa (GERMS-SA).

Supplementary code

Code 2. Read mapping coverage statistics script for generating mean read depth, mapping coverage and creation of read mapping graphs in the region of interest mapping analysis.

```
# IMPORT MODULES
import pandas as pd
import matplotlib.pyplot as plt
import statistics
import optparse
import sys, os

# ARGUMENTS
p = optparse.OptionParser()
p.add_option('--depth', '-d', default=2, action = "store", type="int", dest='d')
p.add_option('-o', default=(str(sys.argv[1])), action = "store_const", dest='out_handle')
(options, arguments) = p.parse_args()

# FUNCTIONS
def findBreadth(b,D,threshold):
    count = 0
    total = len(b)
    for i in range(1,total):
        if D[i]>=threshold:
            count+=1
    breadth =  count/total
    return breadth

# MAIN
df = pd.read_csv(sys.argv[1], sep = "\t", names=["locus","base","depth"])

bases = df.loc[:,'base'].values.squeeze()
depth = df.loc[:,'depth'].values.squeeze()

d = options.d
out_handle = options.out_handle
```

```
out_handle = str(out_handle)+".png"

# Mean Depth
mean = statistics.mean(depth)
mean = repr(mean)
print("mean= " + mean)

# Breadth
breadth = findBreadth(bases,depth,d)
breadth = repr(breadth)
print("breadth = " + breadth)

# coverage diagram
title = str("Read depth, mean depth: " + mean + " breadth: " + breadth)
plt.bar(bases, depth, color = "blue")
plt.xlabel("Genome location", fontsize = 12)
plt.ylabel("Depth", fontsize = 12)
plt.title(title, fontsize = 16)
plt.savefig(out_handle)
```