Molecular Epidemiology and Evolution of *Corynebacterium diphtheriae* and *Vibrio cholerae*

Robert Christopher Will

CITIID

Department of Medicine

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

Word Count: 40,643

Date of Submission: 20/09/2021

Hughes Hall 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text. I further state that no substantial part of my thesis has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. It does not exceed the prescribed word limit for the relevant Degree Committee

Robert Christopher Will

2021

Abstract

Molecular Epidemiology and Evolution of *Corynebacterium diphtheriae* and *Vibrio cholerae*

Robert C. Will

The infectious diseases diphtheria and cholera affect thousands of people around the world every year, and are caused by the bacterial species *Corynebacterium diphtheriae* and *Vibrio cholerae* respectively. While the two pathogens appear quite different biologically, diphtheria and cholera do share key similarities, including the fact that both diseases are intrinsically linked with poverty, outbreaking rapidly in areas of social turmoil, and during both natural and man-made disasters. Both are also highly treatable and survivable with the right resources available, including antibiotics and antitoxins. Vaccines are also available against both causative agents, although the diphtheria toxoid vaccine is in much wider use than cholera vaccine. In this thesis I have used genomics to investigate the epidemiology and evolution of both pathogens in global and national settings. For *C. diphtheriae*, we assembled a large collection of sequenced genomes to investigate the evolutionary and population dynamics of the pathogen across the globe. We added a focus on India as the country with the highest number of reported cases in recent years. We identified multiple large phylogenetic clades of the species circulating across geography and time concurrently. Using this collection, we also identified the presence of antimicrobial resistance determinant genes in recently isolated *C. diphtheriae*. In addition, we categorised a series of variants of the diphtheria toxin gene *tox* from isolates around the world, several of which were non-synonymous and estimated to have an impact on the 3D structure of the toxin protein.

In *V. cholerae*, we investigated the intriguing population and evolutionary dynamics of cholera in Ghana, presenting a picture of time-separated clades circulating around neighbouring countries in West Africa. We also highlight the increasing presence of AMR in West African *V. cholerae*, in line with other reports from the ongoing 7th Pandemic. Finally, we present preliminary analysis of a large *V. cholerae* O139 collection and highlight how rapid AMR development may have caused O139 to so effectively outcompete the existing O1 serogroup, before disappearing almost as quickly due to subsequent loss of AMR.

Taken together, these results highlight how much there is still left to understand about both of diseases, which in many parts of the world are believed to be a problem of the past. This dearth of knowledge applies both to the vastly under-researched diphtheria and the more widely researched cholera.

Acknowledgements

'This is a process. It's a process, it's a process. Okay?'

Moneyball (2011)

I thought long and hard about a quote to include in this thesis. I really wanted to get it right, a set of words that summed up the research I had undertaken somehow, the experience of these three tumultuous years condensed into a few lines. I went back and searched through some of my favourite books and comics, thought about the TV shows and films I enjoyed most, but I was struggling to find one that said what I wanted it to say. In the back of my mind though, I just kept coming back time and time again to this particular quote. Moneyball, the story about the Oakland Athletics attempting to overcome the massive financial disadvantage they have compared to the other bigger teams in American League Baseball. Released in 2011 directed by Bennett Miller and written by Steven Zaillian and Aaron Sorkin, might not seem the obvious choice for a microbiology PhD thesis. Thankfully though, the movie is only partially about baseball, and really is a story of attempting to change a system, how hard progress can be, and how small steps attempted now can lead to much larger leaps taken by others down the line. *'This is a process. It's a process, it's a process. Okay?'* is the line that sums all that up best to me, and perfectly surmises in 11 words one of my core beliefs about science that has only become strengthened during my PhD: that research is a cooperative endeavour, taking steps along the road for others to follow and then advance in their own ways. It's a process, one that I have loved being a part of, and look forward to watching continue long after I have left.

If you had told me seven years ago as I began my undergraduate journey that I would be writing a thesis at the University of Cambridge, under the supervision on two world leading scientists in Gordon Dougan and Ankur Mutreja, I would not have believed you for even a moment. I want to thank Doog for constantly keeping me on track, always having a new idea for me to investigate, and for not being too harsh on the number of revisions we had to go through on my writing! Ankur, I want to thank you for giving so much of your time to me, always being available whenever I had a question or needed help, and for being a soundboard during the difficult times. I am so glad you both took the leap of accepting me into your groups after only a few days of meeting me.

I also want to thank the many people in both the UK and overseas, without whom I could never have completed this research. From Ellen Higginson and Zoe Dyson, for their expert advice and knowledge whenever I needed help, to Josefin Bartholdson Scott, Sally Forrest, Derek Pickard and Fahad Khokhar for bearing with someone with basically no lab experience trying to understand protocols and equipment, and to Ben Warne and Mailis Maes, who atop everything else they did to help me, were going through their own PhD journeys at the same time. Thanks as well to Eva Heinz and Vartul Sangal, for hosting me in Liverpool and Newcastle and giving me crash courses in informatics, teaching me to use many of the techniques utilised throughout this thesis. Agila Kumari Pragasam, Karthick Vasudevan, and Japheth A. Opintan, I am so thankful for your support and amazingly hard work, these projects would not be what they are without you all. Thank you as well to the teams in India and Ghana, for sourcing, growing, and extracting all the novel isolates contained within the genome collections I used, and to Stephen Baker and To Nguyen Thi Nguyen who provided

early access to their collection too. And to Archana Madhav and Billie Pembroke, it has been a great journey alongside you both. I can't wait to see where you both go next. I want to also thank Dan Forman, Naomi Ginnever, Penny Neyland, Elisabeth Cozens, and everyone else at Swansea University who gave me the support, tutorage, opportunities, and references to get to this PhD in the first place. I will always be grateful to have had the opportunity to work alongside so many incredible scientists throughout my academic years. I would also like to thank the Medical Research Council for funding my PhD through the University of Cambridge MRC Doctoral Training Program, Dee Longdin, Suzanne Diston, and Guy Williams for all their help during the PhD process, the University of Cambridge Department of Medicine and CITIID for hosting me, and Hughes Hall for accepting me into the College.

Thanks as well to Eva Scholtus, Charlotte Mitchell, Kirti Mistry, Paul Easton, Peter Cotgreave, and everyone else at the Microbiology Society and ECM Forum who made my membership of the organisation so impactful. It has been great working alongside you all, and the opportunities afforded me where amazing to be a part of and has allowed me to gain so many experiences outside of my research.

On personal notes, I have plenty of people to thank too. Anna, Hannah, Stuart, James, and Jadey, thank you for coming out in all weathers as part of the hedgehog team and for your friendship during our final years at Swansea. To Tim, J.R., Tom, and Henry, long may our adventures across the tabletops continue. Connie, Johnny, Davis, and Alan, you have all been so supportive of the one who went away, and I can't wait to see you all again soon. You are the greatest group of friends I could ever have asked for. To Rachel, I don't think either of us thought that twenty-minute wander from office to office trying to work out how to apply for

placement years would lead to such a strong and lasting friendship, but I am so thankful it did. Steffan, thank you for being my best man, for being our most regular visitor, and for StreetPass-ing me on the very first day we met. I still can't think of a more fitting way for us to have started getting to know each other. And of course, Phil, for being such a constant in my life these past thirteen years. You are my oldest friend, and your constant support and counsel have always seen me through. I am so looking forward to telling stories and rolling dice around your table once again soon!

Finally, thank you to my family for their love and support, and for the times you all tried so hard to understand what on earth I actually spent my days working on. To Kieran, you are an amazing brother, and it has been so wonderful seeing you discover your true calling over the past couple of years. To Nee, thank you for all those times you looked after me, for always being free for a chat on the phone, and for the lunches every time I was home. To Roz and Laura, you are the nicest siblings-in-law anyone could have, and I am so thankful you accepted me into the family. Dad, thank you for all your advice and help, not least of all getting my very nervous self to the PhD interview on time and in the right place. To Mum, you are an inspiration, and the greatest mother I could have asked for. I am eternally grateful for everything you sacrificed for me to be where I am today. And Kim, the fact you agreed to take this wild ride with me is still crazy, let alone agreed to get married right in the middle of it all. You are my rock, and there is no way I would be here today without your love and companionship. Thank you for everything.

Table of Contents

Table of Figures

1.0 Introduction

1.1 History and Evolution of Genomics

1.1.1 Why we need population biology for pathogens

Originally developed as an ecological concept, 'population biology' began as a study of the characteristics and factors that affect the distribution and size of the population of an organism [1]. Bringing together two closely related strands of research in population ecology and population genetics, the field initially exploited mathematical models to understand the interwoven dynamics and selective pressures shaping communities of organisms [2]. These shaping factors included breeding rates, population density, migratory trends and predator-prey organisms. The field built complex formula to attempt to understand how these mechanisms affect each other and the organism or system under observation [2].

In 1979 Roy M. Anderson and Robert M. May wrote a two-part review article highlighting the growing body of evidence that disease-causing microorganisms had a large impact on many of these models by exerting a limit on growth, primarily in place of the predator or resource limitation [3,4]. Since then, many textbooks have included chapters and sections dedicated to the population biology of infectious diseases, often using the example of parasite-host interactions [5]. This has expanded to teaching as well, with many students (including this author) being taught pathogen population examples alongside the traditional pairings of deer and wolves or foxes and rabbits.

In recent decades, the field of study has expanded beyond the original models of interaction, and now encompass epidemiology, pathogen biology and phylogeny to name a few influences. Alongside studying the impact of microorganisms on human, animal, and plant populations, the increasing availability of genetic data has led to vast improvements in our understanding of the structure of microorganism populations, rather than as a factor impacting something else. In particular, vast strides have been made in the understanding of how such microorganisms change over time and space. As genetic analyses began to become the norm, the discipline of population genetics was effectively founded by Ronald Fisher, John Burdon Sanderson Haldane, and Sewall Wright [6,7]. With his 1930 book 'The genetic theory of natural selection' Fisher brought together natural selection as the driver of evolution alongside mathematical analyses [8]. At around the same time, Haldane was applying statistical analyses to determine the frequency change of a gene when placed under specific conditions, and highlighted the now-well-known peppered moth evolutionary example; in areas where pollution levels in their habitats increases and blackened trees, the moths' eponymous black-white speckled colouration changed to a much darker grey-black, suggesting a selective pressure towards the more effective camouflage scheme [6,7]. While both Fisher and Haldane were active in the United Kingdom, in the USA Wright – a biologist with a background in animal breeding – presented mathematical models to better define the 'The distribution of gene frequencies under irreversible mutation' [9]. Some of the implications of this paper would not be realised for three decades, when Motoo Kimura, another pioneer of the field, presented their 'Evolutionary rate at the molecular level' paper in Nature [6,10]. Other leading evolutionary geneticists included William Donald Hamilton, a pioneer of the impact altruism has in genetic fitness, as well as the evolution of sex ratios, and John Maynard Smith, a student under Haldane who, alongside George R. Price, introduced the idea of game theory to

genetic evolution. All three highlighted the importance of understanding the organism and group dynamics, alongside pure genetic analyses [6,11–14].

Genetic variation is at the heart of all phenotypic change. It is critically important to continue understanding how a disease-causing organism are evolving, as this can give researchers a glimpse into how the same microorganism might adapt in the future.

1.1.2 Before genomics

One of the most widely known examples of very early epidemiological studies on infectious diseases is that of John Snow carried out in 1854 [15]. During a cholera outbreak in London, he plotted the location of infected households onto a map using public records and interviews and noted that a single water pump was a potential linking factor [16]. It is now part of folklore that by removing the handle of this pump the cholera outbreak was curtailed. John Snow is now regarded by many as one of the 'heroes of modern epidemiology' [17]. The use of metadata such as the location and time of diagnosis allowed a much higher level of resolution to understand disease spread. This combined with observations on the temporal variances in symptoms or severity of disease allowed practitioners to track the spread from patient to patient across communities.

From the first presentation of symptoms to diagnosis and treatment, understanding the cause of an infection has always been challenging. Understanding the difference between microbial 'strains' (or their phenotype) determined by growing and analysing them in laboratories has been the mainstay method of investigation of outbreaks for decades. This approach has

traditionally relied on the abilities of experts to identify the organism after hours or days of growth and testing [18].

The advent of DNA sequencing however was to provide a quantum leap in our understanding of not only the structure of microbial populations but also the genetic basis of phenotypic variation [18]. One of the first DNA sequencing technologies was developed by Frederick Sanger and his team in 1977 [19]. Known as 'Sanger sequencing', it was used in a 'proof of principle' study to sequence ϕX174, a bacteriophage that infects *Escherichia coli*. ϕX174 has a genome sequence 5,386 nucleotides in length and was the first full genome of a 'life form' to have ever been sequenced. Sanger went on to share the 1980 Nobel Prize for Chemistry for his work [20,21]. Applied Biosystems, a company founded by Leroy Hood, commercialised machines able to perform automated Sanger's sequencing [22]. Their machines, with appropriate technical support, could sequence ~12,000 DNA bases per day, and this breakthrough prompted automated sequencing at an industrial level [23]. Alongside this breakthrough, in 1977 Roger Staden and others proposed using computer programs to assemble genetic reads into sequences [24]. The technologies may have grown and improved far beyond recognition, but the combination of laboratory sequencing and computational assembly is still the basis of all genomic analysis used today.

While the human genome project began in 1990, it took a further five years for bacterial genome sequencing to take a major jump forward. The practice of sequencing individual genes/regions was ongoing, but the field was redefined with the publication of the first complete bacterial genome sequence, an isolate of *Haemophilus influenzae* [22].

1.1.3 The first complete genome (*H. influenzae*)

Despite the advent of consortia aiming to sequence model organisms, it was *H. influenzae* Rd that became the first completely sequenced genome of a bacterium, or indeed any free-living organism [22]. Work was progressing on sequencing ever larger genomes, from bacteria such as *E. coli* and *Bacillus subtilis*, to the fungi *Saccharomyces cerevisiae* (Brewer's Yeast), and on to more complex multi-cellular organisms such as *Caenorhabditis elegans* (a free living worm) and *Drosophila melanogaster* (Common Fruit Fly) [25–29]. Despite all of these consortia expecting to report first, in 1995 Fleischmann *et al.* at the Johns Hopkins University School of Medicine in Baltimore beat them to it by publishing the completed genome of *H. influenzae* Rd [30]. The team presented a method involving sequencing small pieces of unselected *H. influenzae* DNA, that were subsequently assembled into a complete genome [30]. This method is more commonly referred to today as shotgun sequencing, and allowed Fleischmann *et al.* to decode the entire 1,830,137 base pair long genome without the need for an existing genome map – a major limitation of some of the other early attempts to generate whole genome sequences [30].

The strain of *H. influenzae* used was, somewhat ironically, non-pathogenic but still presented an incredible trove of information regarding potential pathogenicity and virulence [30]. The authors noted that the sequenced *Haemophilus influenzae* Rd strain differed significantly from other *H. influenzae* serotype b isolates that causes disease, including the absence of a gene cluster used by the bacteria for adhesion in host cells, that they hypothesised to have been lost by deletion [30].

By using the complete genome to investigate differences between strains, researchers were able to achieve a previously unattainable level of genetic resolution, able to analyse individual gene gaps and additions. With an uncertainty rate of <1.5% coupled with rigorous assembly algorithms, whole genome shotgun sequencing had demonstrated its ability as a method to produce high quality genomes [30]. It was theorised that the potential amount of information available within the 1.8 Mb genome could stimulate leaps in vaccine development as well as other industrial applications [30]. The era of creating reference genomes had begun.

1.1.4 The era of creating reference genomes for different pathogens

In the 26 years since *H. influenzae* Rd was sequenced, there has been a quantum leap in the number of genomes available. Before the advent of next generation methods and large scale mass sequencing however, the first genomes of many important pathogenic species were being published [22]. A mere three months after *H. influenzae* Rd's sequenced was published in Science, the journal once again hosted a genome sequencing announcement and analysis; that of an isolate of *Mycoplasma genitalium*. Less than half the size of *H. Influenzae* at 580,070 base pairs, Fraser *et al.* reported it as the smallest known genome of any free-living organism, with only 470 coding regions identified [31]. The team that sequenced *H. influenzae* Rd was also involved in the *M. genitalium* work, and once again demonstrated the power of the shotgun sequencing method [31,32].

Over the next few years, a number of genome sequencing projects and consortia began to report complete genomes of bacteria [22,32]. Fourteen months after the first *M. genitalium* sequence was published, the *Mycoplasma* genus became the first to have isolates from two different species sequenced completely; *M. genitalium* and *M. pneumoniae* [33]. In September and November 1997 respectively the first *E. coli* and *B. subtilis* genomes were published, while the genomes of *Mycobacterium tuberculosis* and *Rickettsia prowazekii* isolates both followed shortly after in 1998, the latter demonstrating the phenomenon known as reductive evolution [34–37]. The last year of the 20th century saw for the first time a single species, *Helicobacter pylori*, having multiple genomes published [38].

1.1.5 First steps to impact through comparative genomics

The dawn of the 21st century brought in the first steps of translating genomic sequences into evidence and impact. The concept of reverse vaccinology, using computational methods to identify potential antigens for further investigation and vaccine development, was reported through an analysis of the sequence of serogroup B meningococcus (*Neisseria meningitidis*) and comparison with other *Neisseria* genomes in 2000 [39,40]. By comparing the genome of an enterohemorrhagic *E. coli* O157:H7 (EHEC) isolate with the previously sequenced E. coli K-12 genome, Hayashi *et al.* where able to postulate the horizontal transfer of multiple genes linked to virulence in 2001. The authors estimated that as much as 25% of the 5.5 Mb EHEC genome being made up of horizontally-acquired DNA [41]. The genome of *Mycobacterium leprae* was published in 2001, demonstrating a high abundance of inactivated or pseudogenes, an observation of the *R. prowazekii* genome three years prior. This comparison provided further evidence for a reductive evolution in these obligate intracellular pathogens, with less than half the *M. leparae* genome containing functional genes [42]. All these major

breakthroughs came directly from sequencing the complete genomes, and they would not have been made without this step, clearly demonstrating the impact these new methods.

2001 also marked the first time multiple genomes of a single species were published in one manuscript, with two related methicillin-resistant *Staphylococcus aureus* (MRSA) sequenced by Kuroda *et al.* [43]. Comparative genome analysis was used to identify novel antimicrobial resistance determinants in *S. aureus*. As further studies began to include multiple genomes rather than just one, investigations into the genetic differences between isolates and across species started to become more common place. The continual refining and development of new sequencing methods coupled with advancements in sequencing machines meant that the amount of data becoming available was beginning to outstrip the time required to fully analyse it by relying on human methods. The need for computational tools that could take on the challenge of analysing increasingly larger data sets was being made abundantly clear.

1.1.6 Early steps in the informatic analysis of microbial genomes

As the rate of data generation began to outstrip the ability of individuals or teams to perform manual analysis, the development of computation tools to automate parts of these analyses began to accelerate. However, the origins of some of these bioinformatic tools came years before complete genomes were available [44,45]. The true origins of informatics lies in some of the fundamental challenges that have faced biologists throughout history, from elucidating the structure of DNA and the genetic information that encoded proteins to determining the structural properties of protein molecules and the factors that govern them [46–51]. The difficulties that these scientists faced are ones that would come to be part of computational

analysis decades later, as traditional methods were combined with novel digital ways of interrogating data [51]. The comparing of phenotype to genotype, and the takeaways from how well they match, continues to be a key pillar of research to this day.

The combination of computational analysis and biology was embodied by an early pioneer of the field; Margaret Dayhoff [44,45,52]. Referred to as 'the mother and father of bioinformatics' by the former director of the National Center for Biotechnology Information (NCBI) David J. Lipman, Dayhoff was a pioneer of creating and utilising computational tools [44]. A quantum chemist by training, she published an atlas of protein sequences and structures that introduced one-letter codes for amino acids, perhaps the first ever database for molecular biology [53]. These single letter codes are still in use today [54]. It seems fitting then that a pioneer of bioinformatics hails from the same field that spawned the discipline. Some of the earliest computational tools were used to assemble protein sequences and elucidate their structures, even constructing a three-dimensional model by as early as in 1966 (Cyrus Levinthal) [51,55]. While computational hardware was a limiting factor initially, these initial developments paved the way for tools used on a daily basis in 2021.

In 1967 Walter Fitch and Emanuel Margoliash published a seminal paper describing their computational program that could construct phylogenetic trees based on the widely sequenced cytochrome *c* protein. This is a respiratory pigment found in aerobic cells, and the same protein was modelled in three dimensions a year earlier by Cyrus Levinthal [45,55,56]. In 1969, Dayhoff published a similar method that had been developed concurrently, with both methods producing phylogenies that bore high similarity to the traditional trees constructed using taxonomic characteristics [56,57]. This demonstrated a major characteristic of the

bioinformatic field that has continued to today, different groups developing different tools concurrently that use similar but different methods to tackle the same problem.

The difficulties with these early phylogeny-building programs were numerous. The outputs still required human intuition to establish the simplest tree, and while cytochrome *c* was one of the most widely sequenced proteins available, there was still the challenge of determining the difference between homology and chance similarity [45].

Major steps forward followed these early developments, including Needleman and Wunsch's work building upon Fitch and Margoliash's by breaking down construction into a series of smaller steps, Dayhoff, Schwartz and Orcutt publishing the first probabilistic model of amino acid substitutions, and Feng and Doolittle produced a method highlighting the impact of creating progressive sequence alignments prior to correct phylogenetic tree construction, an important prerequisite for accurate phylogeny [44,45,51,58–60]. These approaches proved to be the foundation for modern phylogenetics and continue to be used (in refined forms) today. While new tools continue to be developed, they all owe their existence to these early pioneers [44,45,51]. The first phylogenies were based on small protein sequences, unsurprising bearing in mind DNA sequencing was barely in its infancy as a field [44,45]. Phylogenies became based on multiple larger amino acid sequences, before DNA began to be asserted as the more informative medium for phylogeny construction, with the underlying code providing a higher resolution for any changes that the translated amino acid sequence [44]. As both sequencing technologies and the computational tools to analyse them improved, so did the amount of data available to base trees on. Gene sequences expanded to genome sequences, and very quickly the rate of data production began to outstrip the speed at which it could be analysed

[44]. Today, phylogenies are often limited in the availability of computational resources rather than availability of data, with trees constructed based on whole genome analyses rather than a single gene.

1.2 History of Genomics Population Studies

1.2.1 Different disease examples

While John Snow's water pump analysis is often referred to as the first use of epidemiology, and Fitch and Margoliash's trees the first use of phylogenetics, the first genome-based population study is less well defined. Certainly, by the late 2000s next generation sequencing was beginning to accelerate data generation. The 'Genome Sequencer 20 System' by 454 Life Sciences (later acquired by Roche) released in 2005 was the first sequencer referred to as 'next generation', and Applied Biosciences released the second two years later; the '3500 Genetic Analyzer' [22,61–63]. Institutions such as the Wellcome Sanger Institute (WSI), originally established to sequence the human genome, had moved beyond their original remit , to analyse multiple microbial genomes[64,65].

Described by Loman & Pallen as the 'first genomics super-project', Baker *et al.* sequenced isolates of *Salmonella enterica* serovar Typhi (*S.* Typhi, the cause of human enteric fever or typhoid) with the ambition to investigate the distribution and transmission of the bacteria within an urban district in Jakarta, Indonesia [22,66,67]. They used a human genotyping platform that facilitated high throughput analysis, taking the sequences of novel *S.* Typhi isolates and combining these with reference genomes to produce a collection 143 isolates. These were isolated between 1975 – 2005 [66]. By utilising this combination of phenotypic and genotypic analysis, they were able to identify nine distinct haplotypes (bacteria with group of genes inherited from a single ancestor) that had been circulating for over 30 years within Jakarta.

Figure 1.1 shows the relationship of isolates used within this study, using the then-common single nucleotide polymorphism (SNP) distance method.

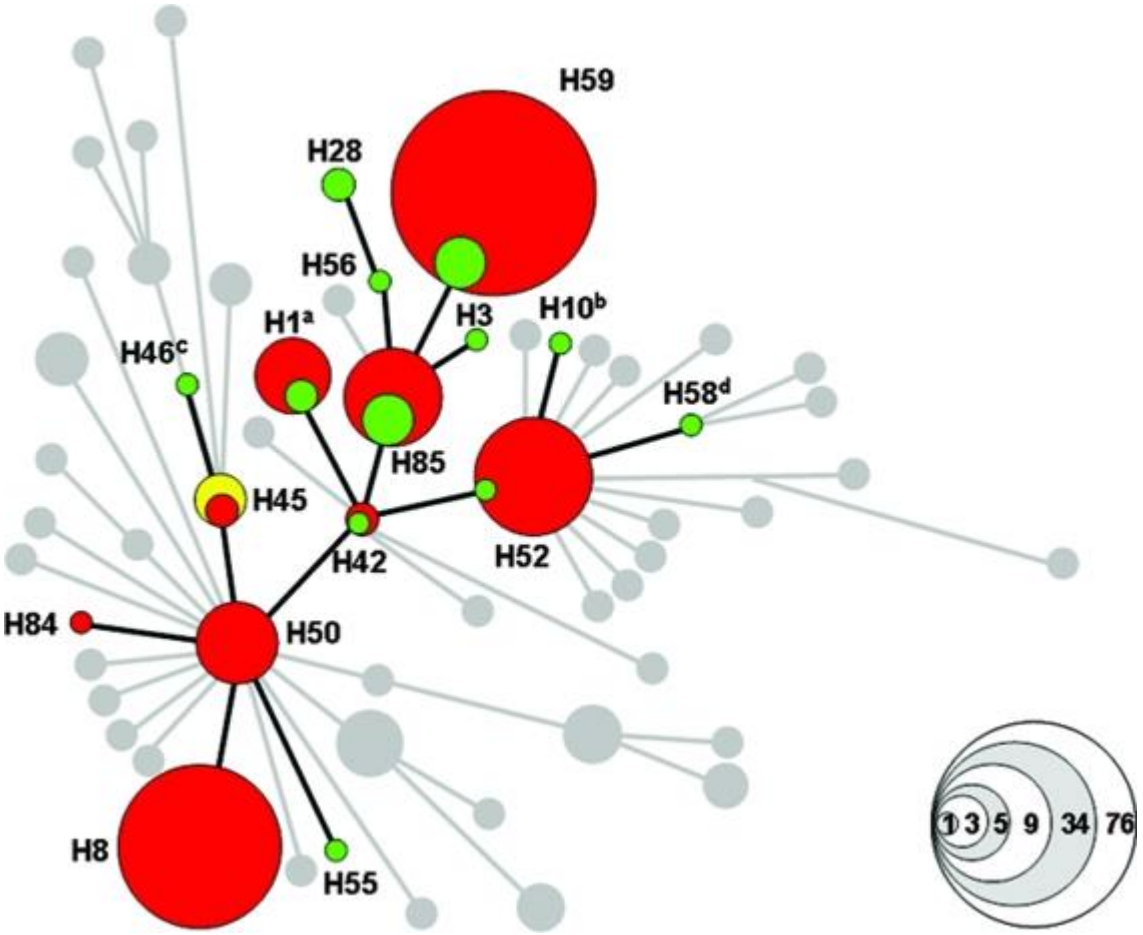

Figure 1.1: *Minimal spanning tree based on single nucleotide polymorphisms (SNPs) showing the relationships between S. Typhi isolates. Taken from Baker et al., 2008, modified from Roumagnac et al., 2006* [66,68].

Alongside the scientific conclusions, this approach demonstrated the power of the high-throughput methodology, combining traditional phenotyping analysis with genotypic tools and genome sequencing [66]. This template quickly became the basis for many landmark genomic studies.

The same team published further analysis of *S.* Typhi, this time using a computational analysis of full genome sequences [69]. Sequencing 19 *S.* Typhi genomes, Holt *et al.* showed that the serovar was associated with ongoing loss of gene function, and despite a rapid acquisition of antimicrobial resistance mutations, *S.* Typhi did not display extensive evidence of antigenic variation [69].

Building and expanding on this method of computational analysis, Holt *et al.* turned their attention to *Shigella sonnei*, a human-adapted cousin of *E. coli* that can cause dysentery through invasion of the human gut mucosa [70]. By whole-genome sequencing 132 *S. sonnei*, the team were able to date a shared ancestor to Europe within the past 500 years, with recent descendants having diversified into a handful of distinct lineages [70]. Thus, the generation of more complex phylogenetic trees based on whole-genome sequences were becoming the norm, bringing together ancestral family trees with increasingly detailed metadata of isolates. An example for *S. sonnei* is shown in Figure 1.2, incorporating country and year of isolation, alongside genetic information.

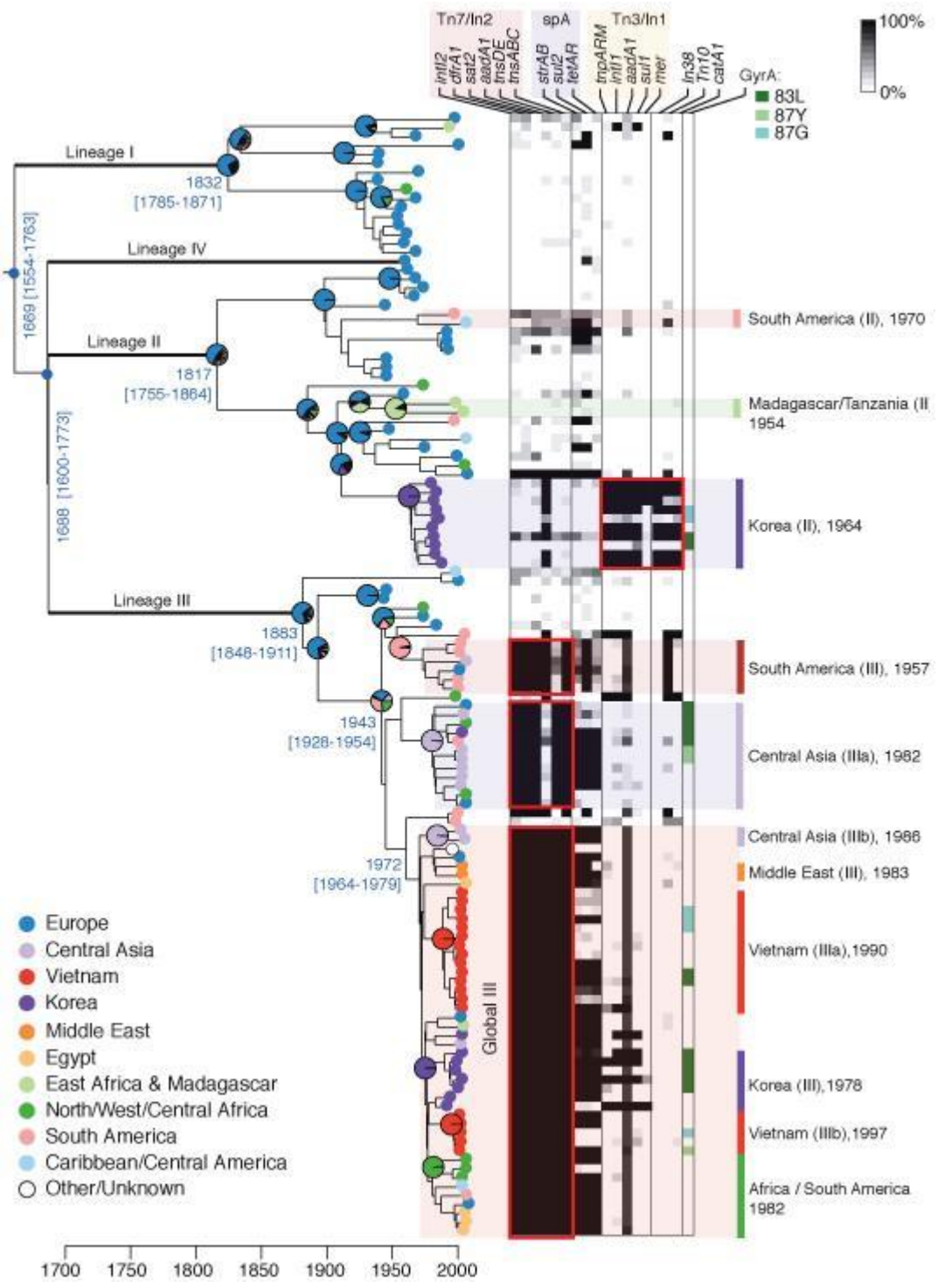

Figure 1.2: *Maximum likelihood phylogeny of S. sonnei, showing the time-adjusted phylogenetic tree alongside key gene presence/absence and metadata showing country or region of isolation. Taken from Holt et al., 2012* [70].

In recent years, the scale and number of studies has continued to increase. In 2015 Holt *et al.* analysed over 328 genomes from animals and humans to determine the diversity, population structure, virulence, and antimicrobial resistance in *Klebsiella pneumoniae*, Connor *et al.* determined the historical global spread and recent local persistence of *Shigella flexneri* using 351 genomes, and in 2019 Van Puyvelde *et al.* created a collection of 276 novel and existing isolates to better understand a single extensively drug-resistant sublineage of *S.* Typhimurium ST313 [71–73]. These are just a few examples but demonstrate the ever-increasing amount of information available in the current era of rapid next-generation sequencing.

1.2.2 Different sequencing technologies

In the years since Sanger sequencing was first invented, many new methods have entered the market. The first 'next generation sequencing' (NGS) method was 454 sequencing, used by the previously mentioned 'Genome Sequencer 20 System' by 454 Life Sciences released in 2005 [19,62,74]. In a process called pyrosequencing (first developed by Nyren, Pettersson and Uhlen in 1993, before years of refinement by Ronaghi *et al.*) single strands of DNA to be sequenced are mixed with enzymes and a nucleotide base, one of adenine, cytosine, guanine, or thymine [75–78]. The binding of a complementary base to the single strand releases pyrophosphates, which are broken down to ATP, which in turn acts as the facilitator for the conversion of luciferin to oxyluciferin. This process releases light bursts which are captured and recorded to determine the next nucleotide base in the single strand. The remainder of unused material is broken down, and the next nucleotide base is added. This process is repeated until the full complementary sequence can be elucidated, before conversion back to determine the original single strand [75–78]. The method offers high accuracy and formed the

foundation upon which high throughput sequencing began, but the approach struggles when determining longer sequences of a single repeated nucleotide and has limited read lengths [22].

The next major sequencing technology to emerge was from Solexa (in what would become known as Illumina sequencing when Solexa technology was acquired). Shankar Balasubramanian and David Klenerman came up with a method coined as 'sequencing by synthesis', stemming from an idea they had in the late 1990s [79]. DNA is purified and adapters added to the single stranded fragment ends, to act as reference and binding points. These are loaded onto flow cells, where fragments are washed across a chip containing oligonucleotides. The fragments bind to these complementary oligonucleotides, which act as an anchor point before cloning begins. Tagged nucleotides and primers are then added, and these nucleotides bind to the sequence one at a time. After each round of one addition, a computer reads which base has been added by the fluorescent tag and all unused elements are washed away, replaced with more for another round of synthesis. This continues across thousands of pieces of DNA, allowing an entire genome to be sequenced in parallel [74,80,81]. Illumina sequencing offered much faster speeds and lower costs, although at the expense of shorter read lengths. Illumina sequencing quickly established itself as a go to method for genome sequencing, a status that persists even today [22,74].

While Illumina remains the mainstay, the aptly termed 'third generation' of genome sequencing technology has begun to establish itself. PacBio Single-Molecule Real-Time (SMRT) sequencing has established itself as an alternative to the traditional Illumina method, by offering much longer read lengths, although at a higher price. Double stranded DNA is broken into fragments and split, before adapters bind to either end and create a circular

sequence. Polymerase and primer are added, and the circular molecule is held within one of millions of wells on a chip. Light is emitted as labelled nucleotides are bound to the DNA sequence, and this is tracked by a computer in real time [82,83].

The other third generation sequencing technology also produces long read data. Called Oxford Nanopore Sequencing, it is a much cheaper method than PacBio to produce long reads, although accuracy is lower. Double stranded DNA is bound with a motor protein and adapter sequence before being washed over a chip containing over 2048 nanopores, by-passing the need for a PCR step. Tethers guide the DNA strands into these nanopores, where the double helix is split, and the forward strand is fed through the nanopore. Electrical current changes characteristically for each base, and the sequence is translated from these variations [74,84,85]. The small size and robustness of the sequencing machines along with its much cheaper price compared to other sequencing methods have created a niche, not only for the laboratory, but in field environments. This was aptly demonstrated during the 2014 – 2016 Ebola outbreak in West Africa, where MinIon sequencers were successfully deployed in remote settings to sequence and investigate Ebola virus genomes in real time [86–88]. The same occurred for Zika virus, also in 2016, when multiple groups utilised the field capabilities of MinIon alongside Illumina after multiplex PCR method for targeted enrichment of the genomes [89].

A newly emerging method of analysis has been the combination of both Illumina short read and either PacBio or Oxford Nanopore long reads to counteract many of the problems associated with each [90–92]. While this remains expensive and requires access to multiple

methods of sequencing, it presents an intriguing future of hybrid assemblies, the most accurately sequenced genomes available today.

1.2.3 Different exemplar computational tools

The core of phylogenetics is the generation of trees. What started with Fitch, Margoliash and Dayhoff has expanded into an incredibly complex and varied field, with numerous methods and tools to create new phylogenies. The workflow however remains relatively unchanged. Raw nucleotide or amino acid sequences, known as Fasta files, are combined into a single file, known as a multifasta file. This multifasta file is then aligned (an important step as stated by Feng and Doolittle) where areas of high similarity between individual sequences are identified and placed alongside each other [93–95]. This allows gaps and areas of difference to be correctly aligned, an essential pre-requisite to all phylogenetic analyses [94]. One of the main tools used for this is Clustal, which aligns the most similar blocks first before working through to the least similar blocks, in a process known as progressive alignment [93,94,96]. Originally created by Higgins and Sharp in 1988, new versions have continued to refine and update the program over the past decades, and has been managed by the same team at University College Dublin to this day [93,94,96–99]. Clustal is so widely used that multiple version publications placed in Nature's top 100 most-cited papers of all time in 2014, at both 10 and 28 on the list [100].

After alignment comes tree creation. While there are many different tools available, two appear more commonly than others; RAxML and IQ-TREE. RAxML has been around much longer than IQ-TREE with version 8 published by Stamatakis in 2014, while IQ-TREE was first published in 2015 by Nguyen *et al.* [101,102]. Both create maximum likelihood phylogenies,

a term based in mathematical statistics which creates trees based on the highest probability of that evolutionary model occurring [103]. The tree-making tool will repeatedly create phylogenies for a pre-determined number of times (known as bootstraps in RAxML, and pseudobootstraps in IQ-TREE) and then choose the structure that has occurred most often to be the most robust final output [101–103].

Even once you have a phylogeny, the tree is not overly useful without metadata to plot on it. The structure of the tree allows you to add information such as location, date and time of isolation, the presence and absence of important genes, or wet laboratory results so as to establish phenotypic patterns. Investigating the presence and absence of genes across an entire genome is however an incredibly large undertaking, let alone across 10s or even 100s of genomes at once as many modern studies require. Tools such as ARIBA and SRST2 are able to screen genomic sequences to look for the presence or absence of dozens of genes from a pre-determined catalogue and report those findings in moments – often in only a minute or two per genome [104,105]. This vastly improves the rate of analysis, although it is reliant on an accurate list of gene sequences prior to running. SRST2 was first published in 2014 by Inouye *et al.*, while ARIBA came three years later in 2017 courtesy of Hunt *et al*. Simon Harris, senior author on ARIBA, also produced a method of running *in silico* Polymerase Chain Reactions (PCR), using primers to search through a sequence and extract sections of it for further analysis [106]. This method has been used by many to extract genes of interest for further analysis.

Once data is acquired, plotting it visually alongside the tree can be undertaken. Command line based users can use tools such as the ggtree package, which builds on the ggplot2

package to create annotated phylogenies through the R software environment [107–111]. For those with a graphical user interface preference, the Interactive Tree Of Life (iTOL) allows for uploading of text files and spreadsheets of information, and will plot those alongside an uploaded phylogeny [112]. Both tools can create high quality scientific figures and allow researchers to analyse their population structure, determining trends over time and space.

1.2.4 Advances in cholera from reference to population studies

The era of genetic sequencing has allowed great strides in the study of *Vibrio cholerae*, the cause of human cholera. Unlike for most bacteria, the genome of *V. cholerae* is split across two chromosomes (this is actually common across the *Vibrio* genus). Consequently, additional challenges are posed for sequencing of the species. Despite this, in 2000 Heidelberg *et al.* published the first genetic sequence of *V. cholerae* El Tor N16961, the genome that continues to be the reference to this day [113]. Chromosome 1 consisted of 2,961,146 bps, while chromosome 2 was 1,072,314 bps long, giving a combined whole genome length of 4,033,460 bps [113]. Figure 1.3 shows the genome circular representation presented by Heidelberg *et al*.

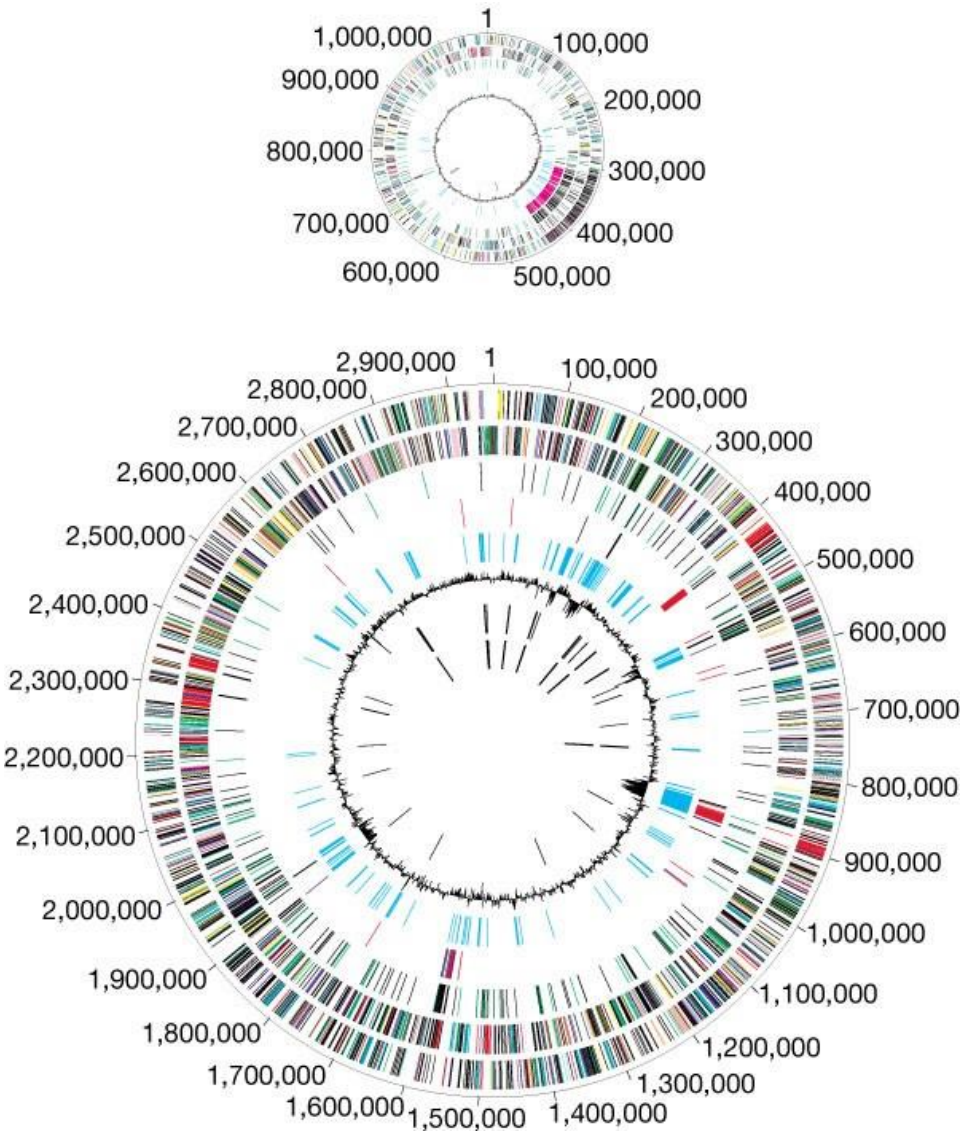

Figure 1.3: *Whole genome circular representation of the Vibrio cholerae two chromosomes. The outside two rings show predicted coding regions, coloured by function, while the third ring shows duplicated genes on each chromosome. The fourth ring shows genes genes relating to phage, pathogenesis, transposons, and VCRs, and the fifth ring shows significant* X^2 *values for trinucleotide composition. The sixth ring shows G+C percentage against the mean G+C content, and the seventh and eighth rings show tRNAs and rRNAs. Taken from Heidelberg et al.* [113].

Between the two chromosomes of N16961, most genes deemed essential for growth are located on chromosome one, including genes responsible for catabolic and biosynthesis mechanisms, and DNA replication and transcription. Virulence-associated genes also follow the same pattern, with significantly more located on chromosome 1 than 2, including the

cholera toxin gene *ctxAB* [113]. The smaller second chromosome is dominated by hypothetical genes, as well as some genes with functions seemingly already catered for by those present on chromosome 1 [113]. Chromosome 2 may be a 'megaplasmid' acquired at some point, with genes from other bacterial species being incorporated into the *V. cholerae* genome at some point in the evolutionary origin of this species [113–116].

Building on the initial reference genome, *V. cholerae* genomic studies began in earnest. Cholera pandemics occur in waves, with the current one known as the seventh pandemic. Chun *et al.* in 2009 used comparative genomic analysis of 23 *V. cholerae* isolates, analysing the basis of genome variation between and within waves [117]. They found that between waves there was a hard 'shift' from one clonal lineage to another distantly related lineage, while variation within a wave was better defined as a slower 'drift', mainly through the addition or loss of gene clusters [117]. A year later, Chin *et al.* used genomics to investigate the then-recent cholera outbreak in Haiti, sequencing five new isolates and combining them with 23 previously published genomes to act as representatives from previous outbreaks [118]. Their results showed that Haitian isolates were closely related to genomes from Bangladeshi cholera cases in the early and late 2000s, rather than those from more geographically close South American isolates [118]. This study, relatively small by modern standards, nonetheless provided evidence for a cholera introduction into Haiti from far across the globe, rather than from more local regions [118]. This presented important policy implications, as the introduction of a potentially higher fitness strain of cholera into Haiti, and potentially beyond, posed a significant risk to human health [118]. The introduction of an antibiotic resistant *V. cholerae* encoding a classical cholera toxin and other novel genes into the Latin American region caused many concerns. These included a potential to replace or recombine with *V. cholerae* already reported in the region[118]. Additionally, as cholera can persist in the environment, once

a new strain has entered the ecosystem, it can be incredibly difficult, if not impossible, to eradicate fully [119].

It was research like Chin *et al.*'s and Chun *et al.*'s that provided the foundation for the first large scale global analysis of *V. cholerae*. Published in 2011 by Mutreja *et al.*, 154 whole genome sequences were combined into a collection, to determine the global community structure, as well as the source of the seventh pandemic [120]. Figures 1.4 shows the annotated maximum likelihood phylogeny of global *V. cholerae*, and Figure 1.5 shows the transmission events inferred from it.

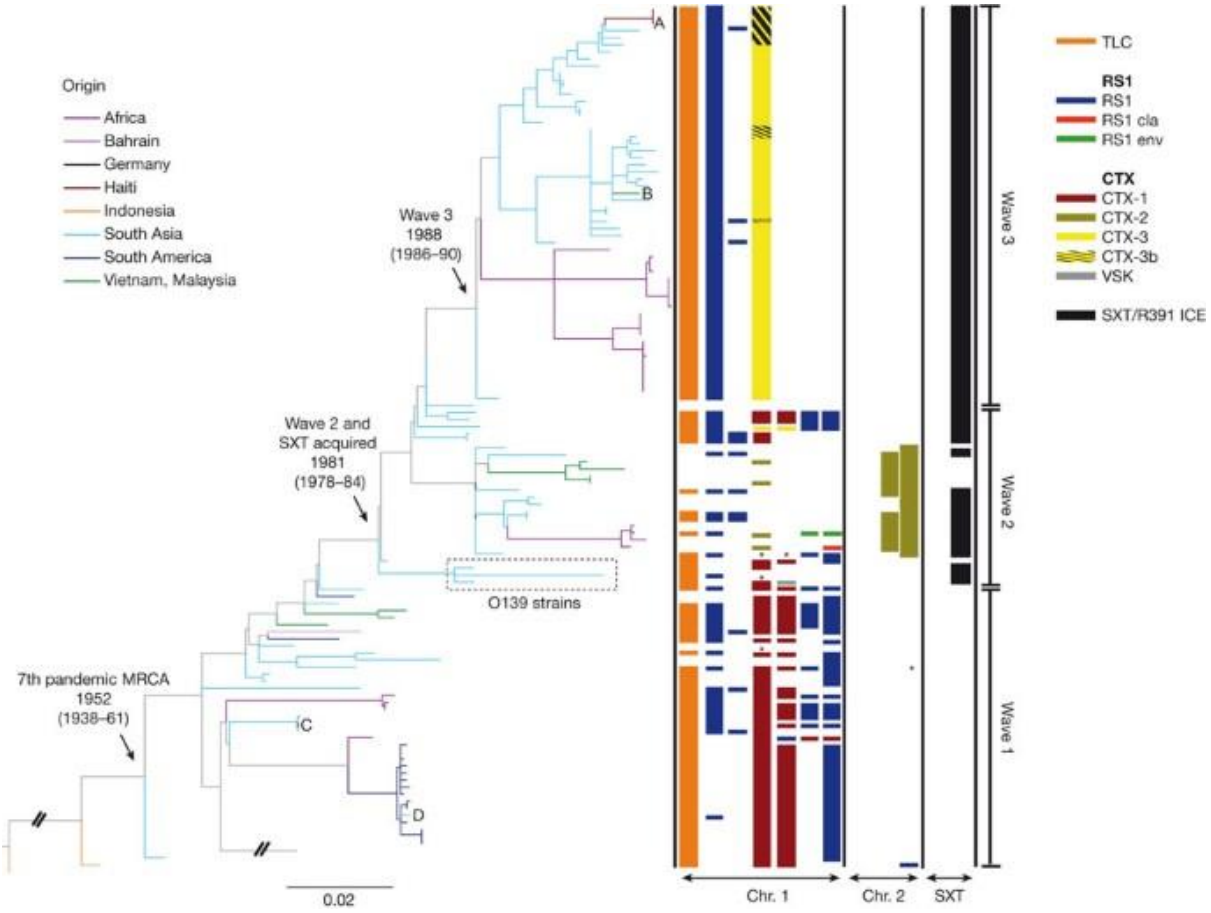

Figure 1.4: *Maximum likelihood phylogeny of whole genome V. cholerae genomes. Metadata is annotated, including the country or region of isolation, the presence of key genetic elements, and the transmission waves. Taken from Mutreja et al., 2011* [120].

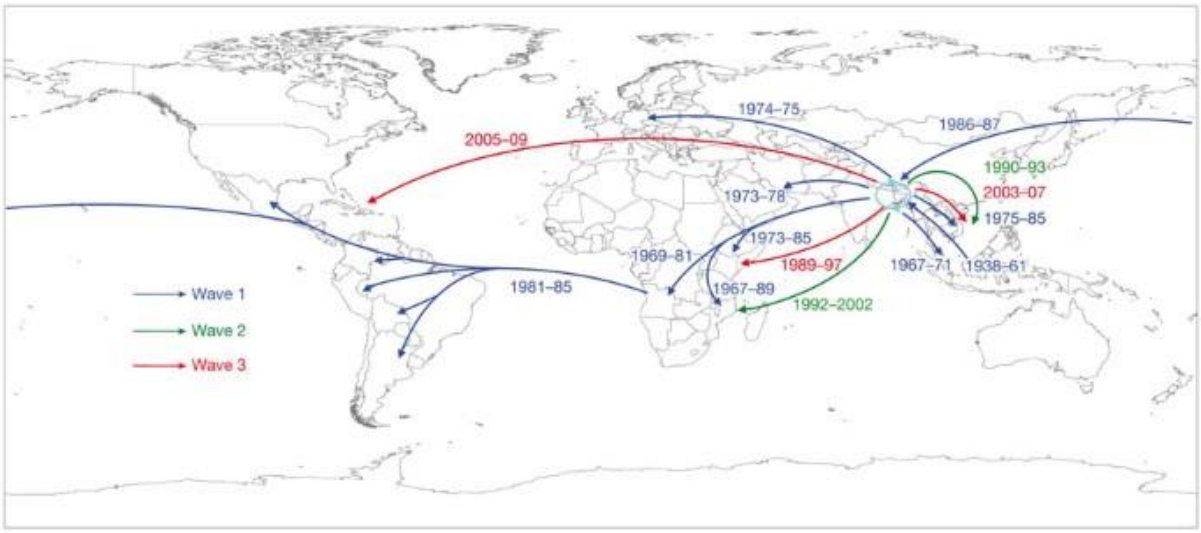

Figure 1.5: Predicted waves of transmission events of V. cholerae, determined from seventh-pandemic time adjusted phylogeny. Taken from Mutreja et al., 2011 [120].

By creating a time-scaled phylogeny, Mutreja *et al.* were able to present a three-wave model for the seventh pandemic, starting in the Bay of Bengal before spreading out across the globe [120]. Over time, these were replaced in some areas by wave 2 and 3 by transcontinental transmission events, a phylogenetic observation that was confirmed by clinical cases [120]. Antimicrobial resistance (AMR) was shown to have shaped these waves, with the development of resistance predicted to have occurred around 15 years after initial drug exposure [120]. This study was a landmark moment in the field of cholera genomics, and the pandemic wave terminology defined by Mutreja *et al.* continues to be definitive today.

In the years since, further studies have expanded on the work of Mutreja *et al.*, both in terms of single outbreaks and transfers across continents. Shah *et al* explored the epidemiology of a 2010 cholera outbreak in Pakistan using genomics [121]. The team found that there were two distinct subclades present across Pakistan (PSC-1 and PSC-2), both within the third

transmission wave of the seventh pandemic [121]. PSC-1 was found to have originated from coastal regions, while PSC-2 was traced back to areas further inland that had been flooded by the Indus River [121]. One of the largest rivers in the world, the Indus flows from the foothills of the Himalayas in Southwestern Tibet through almost the entirety of Pakistan to the Arabian Sea Southeast of Karachi [122]. Understanding the sources of outbreaks is extremely important, and Shah *et al* demonstrated the power of genomics in determining these factors.

Domman *et al* and Weill *et al* published an 'Integrated view of *Vibrio cholerae* in the Americas' and the 'Genomic history of the seventh pandemic of cholera in Africa' respectively in the same Science issue released in November 2017 [123,124]. Both used whole-genome sequencing to categorise novel isolates, before combining the data with additional genomes from publicly available sources to determine a comprehensive picture across both continents, as well as their context within global pandemic transmission events and waves over 40 -50 years [123,124]. Both studies indicated that intercontinental transmissions brought *V. cholerae* to these continents in the seventh pandemic.

In 2019, Weill *et al.* published an investigation determining the phylogenetic relationships, AMR, and virulence determinants present in *V. cholerae* genomes isolated in the 2016 - 2017 cholera outbreak in Yemen [125]. They found that a single sublineage of the seventh pandemic was responsible for both waves of the epidemic, and that it originated is South Asia before moving through East Africa on its way to Yemen [125].

These studies show the massive benefits of having a well-defined global phylogenetic structure of a pathogenic species, allowing future work to understand how their novel isolates fit into the global picture, and determine their potential source and transmission routes.

1.2.5 A different picture for diphtheria

Although we have a good understanding of the global phylogenetics *V. cholerae,* the same cannot be said about *Corynebacterium diphtheriae*. The first genome of *C. diphtheriae* (NCTC 13129) was published in 2003 by Cerdeño-Tárraga *et al.* [126]. The bacterium was isolated in the UK in 1997 from a 72-year-old female returning from a Baltic cruise, and was found to be 2,488,635 bp long [126]. Figure 6 shows the circular representation of the chromosome.

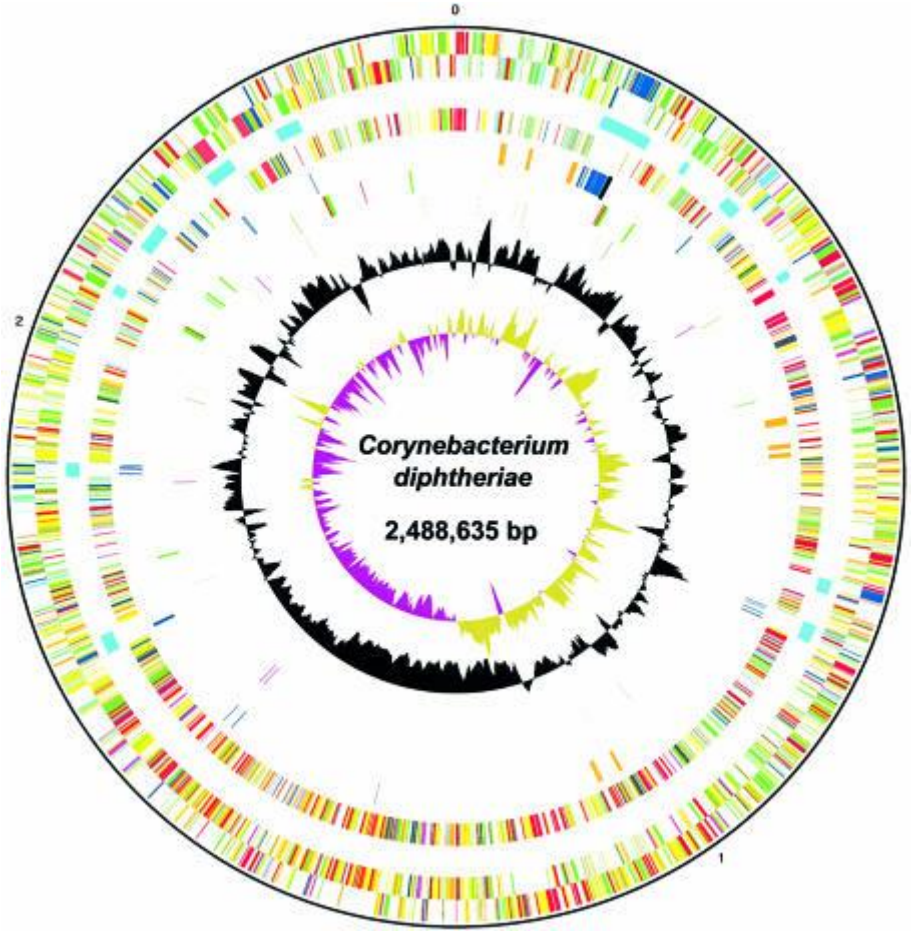

Figure 1.6: *The circular representation of the first C. diphtheriae genome sequenced; NCTC 13129. The outermost ring show DNA bases, with genes presented on the second and third rings. The presence of pathogenicity islands are shown the fourth ring, and gene analogous in Mycobacterium tuberculosis are presented on the fifth ring. The sixth ring shows key genes related to metal-ion transport systems, phages, diphtheria toxin, while the seventh ring shows putative sortases and sortase substrates, The eighth ring shows repX and IS element pairs, and the ninth and tenth G+C content and GC skew respectively. Taken from Cerdeño-Tárraga et al., 2003* [126].

Unlike the related species *Mycobacterium tuberculosis*, the genomic sequence of *C. diphtheriae* demonstrated evidence of recent horizontal gene transfer, including the toxin gene *tox* that is so stereotypical for the species [126]. Recombination is a key driver of diversity across the species, both historically and in modern time [127,128]. Cerdeño-Tárraga *et al.* also determined the genomic location of the Corynephage, along with annotating numerous genes, domains and pathways using predictor tools and by hand [126].

The next step along the path to widespread genomic studies of *C. diphtheriae* was taken by Trost *et al.* in 2012, when they published the first pangenomic analysis of the species [129]. Sequencing and annotating 12 isolates from patients, they combined these with the reference NCTC 13129 to present the first comparative study at scale for *C. diphtheriae*. Once again, evidence of horizontal gene transfer was found [129]. Phylogenies of the 13 isolates were presented based on both allelic profiles of housekeeping genes and the core genes [129]. Due to high levels of recombination, mapping based approaches traditionally used to construct phylogenetic trees were less effective, and a core gene approach provides an alternative. Genes present in 99% of annotated genomes were concatenated together to produce an alignment, limiting the effect of areas with high diversity that could skew trees construction incorrectly.

There are numerous reasons why *C. diphtheriae* lags behind *V. cholerae* in terms of both sequenced genomes and genomic studies. One of the most obvious is the lower number of diphtheria cases reported compared to cholera. Another is the smaller scale of outbreaks. Since the introduction of the vaccine in the 1930s – 1940s, diphtheria has remained largely contained in countries with high vaccine coverage [126]. Despite this, there have been many outbreaks in areas of low vaccine coverage, and while most have been smaller in scale an exemplar large post-vaccine diphtheria outbreak occurred during the dissolution of the Soviet Union in the mid-1990s [130]. Published in 2017, Grosse-Kock *et al.* used the sequenced genomes of 93 *C. diphtheriae* isolates collected from 1996 – 2014 in Belarus, a former Soviet State, alongside previously sequenced representatives [130]. Two major *C. diphtheriae* clones were identified accounting for 76% of the isolates, with47% of the 93 genomes harbouring

the toxin gene [130]. Recombination was once again identified as a key driver of diversity across the genomes [130]. Consequently, a core gene approach was once again used, as had been done by Trost *et al.* [129]. Figure 1.7 shows the core gene phylogenetic tree of genomes isolated from Belarus alongside global representatives.

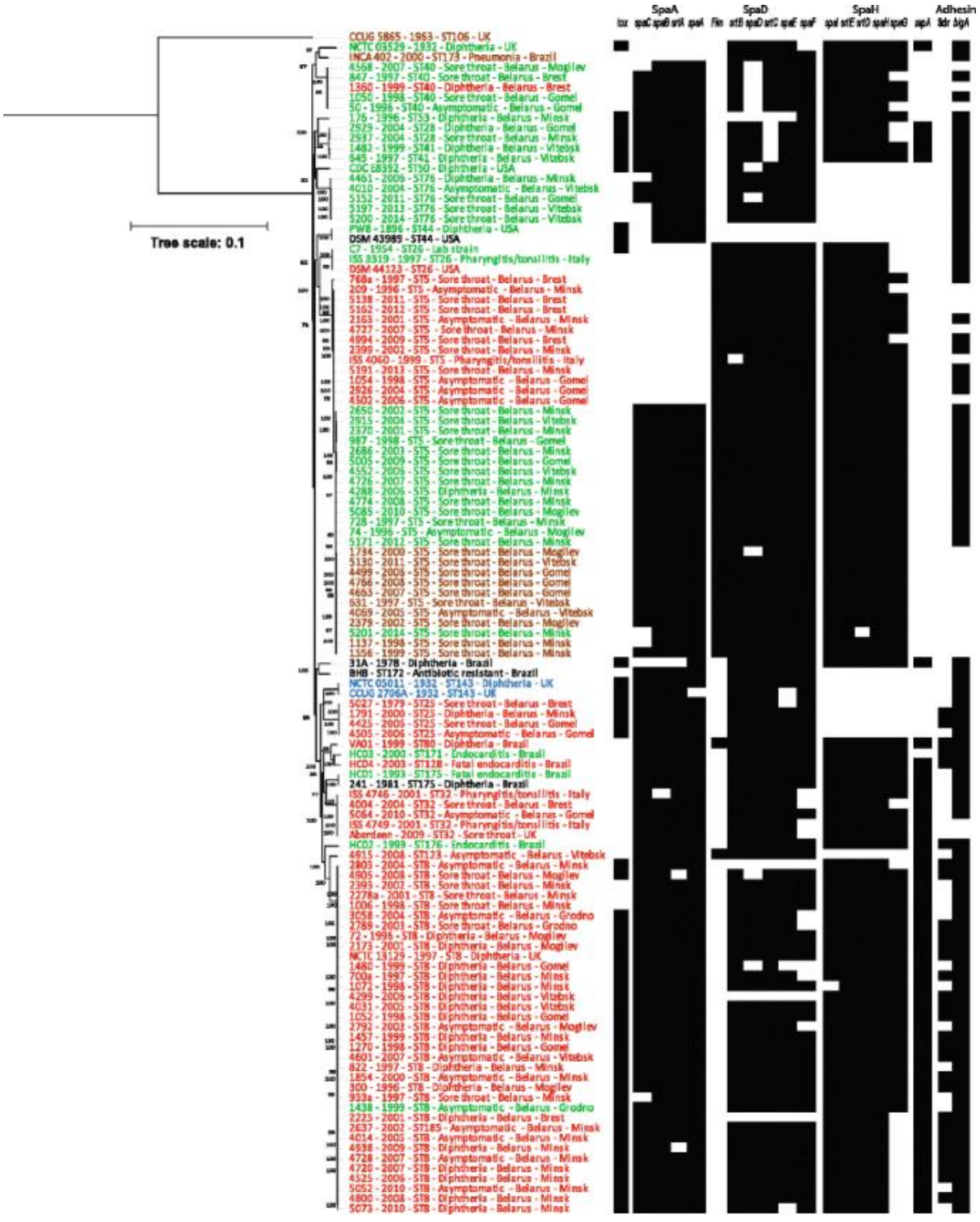

Figure 1.7: *Maximum likelihood core gene phylogeny of Belarusian C. diphtheriae genomes and previously published representatives. Isolates are coloured by biovar designation, and the presence and absence of virulence gene clusters are shown in black and white. Taken from Grosse-Kock et al. 2017* [130].

Outside of this paper, most studies on *C. diphtheriae* genomics have been published describing isolated outbreaks. Examples include papers by Lodeiro-Colatosti *et al.* and Dangel *et al.,* both in the same 2018 issue of Emerging Infectious Diseases [131,132]. Lodeiro-Colatosti *et al.* relied on traditional microbiological techniques, similarly to other studies that came before [131,133–135]. Dangel *et al.* used core genome Multilocus Sequence Types (MLST) and whole genome SNPs to determine the phylogenetic relationship between geographically diverse clusters of nontoxigenic *C. diphtheriae* isolated in Germany [132]. This method of core genome MLST was once again used by Chorlton *et al* in 2020 to support their core gene SNP alignment phylogeny of 56 inner-city Vancouver cases [136]. Timms *et al.* also used a core gene phylogeny to investigate the diversity of 48 *C. diphtheriae* isolates from a 12 year period across Australia [137]. All of these publications focused on diphtheria outbreaks that would be considered very small by cholera standards, keeping the scale small to only encompass outbreak isolates and some reference genomes (although in some cases no references were used) [132,136,137]. The aim of these have been to evaluate *C. diphtheriae* diversity at a small scale, rather than to establish any regional or global strcuture.

The largest study to date has been from Hennart *et al.* in 2020, encompassing 163 isolates from the French mainland and overseas territories combined with 84 historical and reference isolates [138]. While the main focus of the paper was determining thr structure and diversity of a novel AMR plasmid, they too used a core gene approach to phylogeny creation (although using 95% as the cutoff for core definition). They placed novel isolates alongside the representatives to understand the diversity within the French territories and across the collection [138].

The global understanding and framework for the population structure of *C. diphtheriae* would be valuable for multiple reasons [139]. Indeed, as more genomes are published, a comprehensive global analysis would provide an ever-improving understanding of the *C. diphtheriae* population, as the field plays catch up with other pathogens. This is especially true for areas of the globe where there are no or only a few published genomes, as evolutionary changes occurring here are being completely unobserved genomically.

1.3 A History of Cholera

1.3.1 The bacterium

A Gram-negative comma shaped bacterium ~1.3 µm in length, *V. cholerae* has been infecting humans for potentially thousands of years [140,141]. Indeed, ancients texts written in India by Sushruta Samhita in the 5th century B.C., and Hippocrates and Aretaeus of Cappadocia in Greece during the 4th and 1st centuries A.D. speak of cholera-like symptoms, as do Persian writings by Rhazes and Avicenna in the 10th and 11th centuries A.D. [140]. While it was Robert Koch who raised the profile of *V. cholerae* and was for a time credited with the discovery, it was in fact Filippo Pacini that first isolated the comma-shaped bacterium and declared it as the cause of cholera disease [142–144]. It was not until years later in 1884 though that a pure culture was isolated [140].

V. cholerae infects humans through contamination of food or water, colonising the intestine after surviving passage through the digestive system [145,146]. Until the early 1980s it was believed that the relatively anaerobic and motile *V. cholerae* could not survive outside of the human body, and that repeated reinfection of the water systems was required to sustain populations of the species outside of humans [146,147]. Since then however, it has been demonstrated that the *V. cholerae* species is native to aquatic environments, persisting without human influence [147,148]. It has been reported that *V. cholerae* has an average swimming velocity of 75.4 +/- 9.4 microns/sec [141].

There have been more than 200 serogroups of V. *cholerae* defined, all based on structural differences of the O-antigen [149]. Despite this, major pandemics and epidemics in modern history have been caused by only two – the O1 and O139 serogroups. Figure 1.8 shows a timeline of key broad *V. cholerae* serogroup evolutionary changes, taken from Safa *et al.* [149].

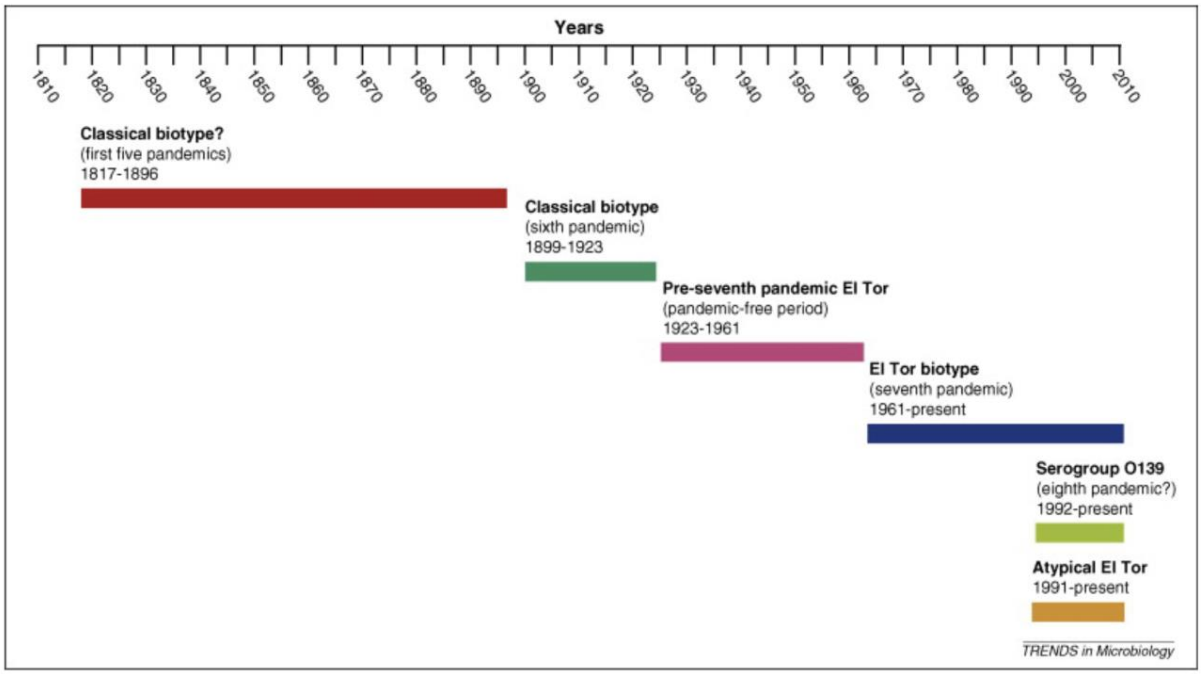

Figure 1.8: *Timeline of key V. cholerae serogroup evolutionary changes. Taken from Safa et al., 2010* [149].

Most isolates from both the O1 and O139 *V. cholerae* biotypes carry the prophage CTX, which encodes the cholera toxin gene, while most isolates from other groups are phage negative [149]. Within O1, which has been the dominant serogroup for at least the sixth and seventh pandemics, the El Tor biotype variant rose to become the driving force of the seventh pandemic [150]. El Tor was initially proposed as a separate species (*Vibrio El Tor*), before being defined as a variant of the O1 serogroup, and it has been shown to have longer survival times on food and within water both in laboratory and field settings [151]. *V. cholerae* O139 on the

other hand was first reported in 1992 during an epidemic across India and Bangladesh [152,153]. Symptomatically, O139 is indistinguishable from O1 cases, producing large amounts of cholera toxin and requiring similar treatment [153]. *V. cholerae* O139 demonstrated AMR comparable to some later O1 isolates, and multi-drug resistant (MDR) isolates were reported in 1996 [145]. *V. cholerae* O139 was even suggested as the start of an eight pandemic due to it outcompeting and displacing existing O1 isolates in some local environments [154,155]. Despite these seemingly strong advantages, O139 outbreaks were large but sporadic, and the serogroup seemingly diminished in frequency post 2005, with only small outbreaks reported in the years after [156].

1.3.2 Vaccines

An effective vaccine against *V. cholerae* infection has long been proposed as a method to control the spread of the disease while offering immunity to those most at risk. Killed whole-cell oral cholera vaccines have been available for three decades [157,158]. Despite this, Vietnam is the only country that incorporates cholera vaccination into its national control program with a locally approved vaccine [157]. Many studies have been published purporting new formulations, with testing occurring in numerous at risk communities ranging from Indonesia, Vietnam and Bangladesh to Zambia, Malawi and South Sudan over the past 24 years [159–165]. While it remains an incredibly well researched area, the World Health Organisation's (WHO) position paper in 2017 reaffirmed their belief that vaccines are only one component of cholera prevention and control strategies, and should not obstruct the continuing of current methods of control [166]. These methods include expert management of cases and outbreaks and continued constant surveillance alongside water-sanitation-and-hand-hygiene (WaSH) interventions, and crucially, education [166].

This position is driven by the fact that while the protection level starts high (>80%), currently approved vaccines demonstrate a decrease in effectiveness to ~60% or less within three - five years, and a shorter duration of protection in children [167]. National agencies in the USA and UK suggest travellers should be vaccinated if visiting high risk areas [168,169]. The development of new improved cholera vaccines remains a high value area, with major funders like the Bill and Melinda Gates Foundation including it in their enteric and diarrheal diseases strategy overview and funding ongoing work in the area [163,170].

1.3.3 Disease and diagnosis

Symptomatically, cholera is defined by 'rice water stool' production with large amount of very watery diarrhoea that can lead to severe dehydration and death within 24 hours if left untreated [171]. Traditionally caused by the release of the AB cholera toxin (CTX), untreated case fatality ratios can be as high as 30 – 50% [171]. Other symptoms can include vomiting and cramps, with an estimated 1 in 10 experiencing severe symptoms [171,172]. While treatment is usually confined to oral rehydration therapy (ORT) and rest, antibiotic treatment is also used in some cases, and these include doxycycline, azithromycin, and tetracycline [173].

Diagnosis technologies have come a long way in recent years, but for *V. cholerae* live culture remains the gold standard for diagnostics [171]. This present numerous challenges, as even excluding the requirement for laboratory equipment and the expertise to operate effectively, there is still the 24-hour growth period to contend with, as well as any transport duration and crucially human error [171]. Time is very much of the essence when diagnosing disease, and these delays can cause serious and life threatening complications [171].

The necessity for speed has led to the development of rapid diagnostic tests (RDTs) for *V. cholerae*. Rapid chromatographic-immuno assays (CIAs) and polymerase chain reaction (PCR) are common techniques utilised by RDTs, due to their potential high specificity and speed compared to the traditional culture method [171,174–176]. CIAs use antibodies designed against the target to bind antigens from a sample on a membrane strip, while PCR uses DNA primers that bind to target sections of a genome present within the sample [171,174–176]. The ability to be deployed into field environments at low cost make RDTs the natural step

forward, as delays in diagnosis, especially in remote locations, quickly leads to further cases and deaths [177]. Additionally, as the robustness of these tests continues to improve, the concern that increases in speed must be bought with a decrease in accuracy becomes less and less relevant [175].

1.3.4 Cholera epidemiology

While the Bay of Bengal is believed to be the source for pandemics of *V. cholerae*, the disease has successfully spread to many parts of the globe. Over the seven pandemics, multiple waves of transmission have spread cholera across continents, conveyed by human travel and contamination (Figure 1.5) [120]. The risk of transmission across and through water bodies, including lakes and rivers, poses a massive challenge for control, as does *V. cholerae*'s ability to remain viable through long periods within those environment [148,178].

Cholera cases are reported by many countries every year, and are collated by the WHO [179,180]. Figure 1.9 shows the case numbers from 1989 – 2017 coloured by region, as published by the WHO [180]. Despite Asia being the ancestral home of *V. cholerae*, African nations have been reporting the highest number of cases for the past couple of decades, passing The Americas in terms of numbers in the mid-1990s [180]. Through 2007 – 2011, 20 nations within the continent reported case numbers of over 100,000, and fatality rates of 2 – 5% [181].

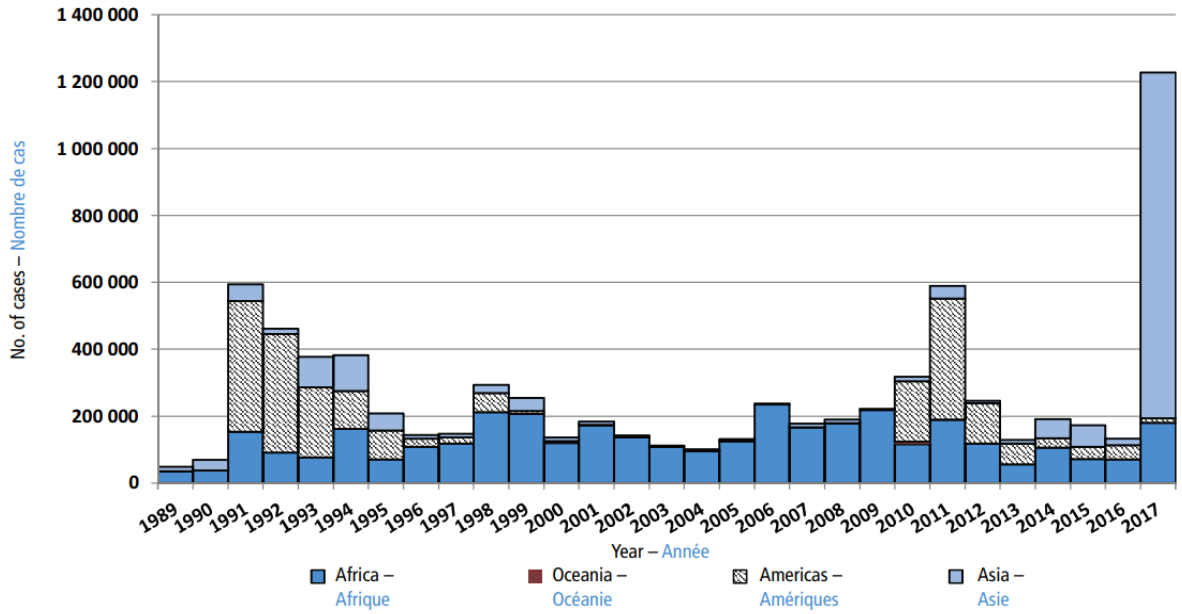

Figure 1.9: *Cholera case numbers from 1989 – 2017 reported to the WHO. Taken from World Health Organisation, 2018* [180].

It was not until 2017 that Asia once again reclaimed the dubious honour of highest number of cases, mainly driven by an explosion of cases in Yemen [182,183]. Figure 1.10 shows the major cholera outbreaks that year. Over a million cases were reported in Yemen, making it one of the largest disease outbreaks in modern history [182,183]. A country with a population of ~25 million, the scale of this outbreak was colossal, and a fatality rate of 0.22% caused 2,385 deaths by March 2018 [183]. Already a poor country in the region, Yemen had been devastated by a war since March 2015, displacing over 3 million people [183]. Cholera cases quickly sprang up as healthcare infrastructure and systems collapsed, and seasons of drought followed by flood exacerbated the spread of contamination [182]. Becoming an unfortunate case study demonstrating the importance of cholera surveillance and control, the outbreak took years to manage, requiring an international effort from many countries and Non-Government

Organisations, highlighting the importance of international cooperation in dealing with global health crises [184].

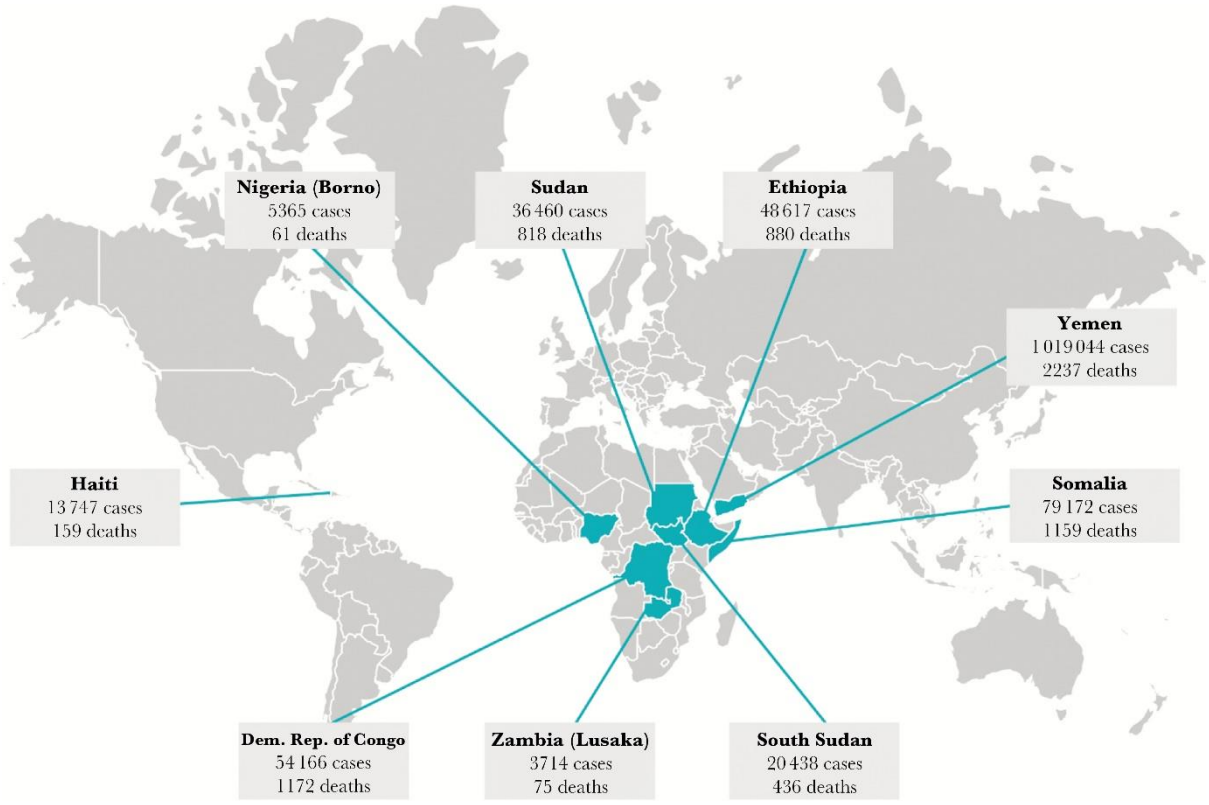

Figure 1.10: *Major cholera outbreaks in 2017, the highest year for case numbers in 28 years. Taken from Legros, 2018* [185].

It is widely believed that official cholera case numbers are underestimates, in part due to the under-diagnosis of the disease in remote environments [186,187]. The further development and deployment of RDTs to these areas will aid in improving diagnostic coverage, and present a more accurate picture of the global burden of this majorly important but preventable disease [186,187].

1.3.5 Impact of genomics on our understanding of cholera

Genomics has improved our understanding of *V. cholerae* on a global level, achieving a degree of resolution rarely matched in the field of pathogen population biology. Building upon the knowledge that there have been seven pandemics of cholera, Mutreja *et al.*'s publication in 2011 provided the framework, showing how the seventh pandemic had been shaped by three waves of intercontinental transmission, upon which many large-scale future studies have been based [120]. This shared framework has allowed for a continually expanding global picture, using a shared structure and language that many other pathogens lack.

Genome sequencing allows for a much higher resolution of detail that previous methods, including MLST. By analysing whole genomes, additional information can be derived, such as the historical origins and potential transmission routes of an individual outbreak [188–190]. Understanding this has implications beyond just scientific knowledge generation, with interesting takeaways for many fields from history to social science and policy [188].

Genomics is also playing a part in improving our understanding of the ecological niche occupied by *V. cholerae* [191,192]. By utilising whole genome sequences genes, SNPs, proteins, plasmids and phages can be found, providing many new avenues for analysing diversity and biology [191]. In the ever-evolving world of bacterial pathogens, it is imperative to not only gain a current understanding of the species but to model how *V. cholerae* might evolve in the future.

1.4 A History of Diphtheria

1.4.1 The Bacterium

A Gram-positive bacterium that colonises the lower throat, *C. diphtheriae* is a very different species to *V. cholerae* [128]. While *V. cholerae* is highly motile, *C. diphtheriae* remains non-motile [128,141]. *C. diphtheriae* is also an aerobic species, the environment of the lower throat differing vastly to that of the small intestine [128]. *C. diphtheriae* is the larger of the two bacteria, at ~2 µm in length compared to *V. cholerae*'s ~1.3 µm [128,141].

Despite these differences, both *V. cholerae* and *C. diphtheriae* do share some important similarities. Both species do not have a spore phase, and both traditionally cause disease by secreting an AB toxin, both of which are coded for by bacteriophage (In diphtheria, diphtheria toxin (DT) is releases in low iron environments) [113,128,193]. Crucially from a global health point of view, both diseases can be controlled relatively successfully with the right infrastructure and systems in place but can quickly cause outbreak if conditions are right.

C. diphtheriae was first demonstrated as the causative agent of the respiratory disease diphtheria in 1883 by Edwin Klebs, and a year later the bacterium was cultured for the first time by Friedrich Löffler, as one of the first proofs of Koch's Postulates [128,194,195]. In 1888, Roux and Yersin presented evidence that it was products produced by *C. diphtheriae*, rather than just the bacteria itself, that caused disease symptoms in diphtheria cases, determined by injecting animals with sterile filtrates of liquid *C. diphtheriae* cultures [196]. It was not until a

vaccine was developed and release 4 - 5 decades later in the 1920s and 30s that diphtheria cases started to drop worldwide, ceasing to be a leading cause of childhood mortality [128].

Transmission of the bacterium occurs by the inhalation of contaminated water droplets in the air, or by contact with a contaminated surface [193]. Unlike *V. cholerae*, it is not believed *C. diphtheriae* can survive long in external environments, and it has only been sporadically found in animals, the majority of which have had easily-traced close contact with infected humans [197].

Finally, 4 biovars (variant strains based on physiological differences) have been reported in *C. diphtheriae*; gravis, mitis, intermedius, and belfanti [198]. While traditionally included in the diagnosis step, whole genome sequencing analysis has suggested that there is no consistent genetic basis for biovar differentiation [198]. Others have proposed that biovar belfanti should be considered a separate *Corynebacterium* species [199]. Outside of *C. diphtheriae*, other *Corynebacterium* species have been shown to cause diphtheria-like infections, most notably *Corynebacterium ulcerans* and *Corynebacterium pseudotuberculosis* [200–203]

1.4.2 Disease and Diagnosis

The definitive symptom of diphtheria is the pseudomembrane; a white-grey build-up of dead cells over the pharynx, larynx and tonsils [193]. Alongside this, diphtheria usually causes severe flu-like symptoms, including angina, sore throat, and mild fever, as well as ‘bull neck’, a swelling of the lymph nodes, although any of these symptoms may be absent in some clinical cases [127,128,193]. The disease has historically been caused by toxigenic *C. diphtheriae*,

secreting diphtheria toxin encoded by the *tox* gene, itself carried on a corynephage [204]. In recent decades, non-toxigenic diphtheria has come into focus. Those *C. diphtheriae* isolates that either do not carry the phage, or whose *tox* gene has become defunct (known as non-toxigenic toxin-bearing, or NTTB), have still shown the capability to cause invasive or systemic infections, although the exact pathogenesis of which are yet to be determined [135,205–207]. Diphtheria case fatality rates remain higher than many other treatable and controllable diseases, exceeding 10% in some areas [208].

The gold standard for diphtheria diagnosis is, much like cholera, culturing from a clinical specimen, although in diphtheria this is followed by a toxigenicity test known as an Elek Test [209]. This presents the same challenges as with cholera, where time is often of the essence when treating diphtheria, but diagnosis is often delayed for well over 24 hours due to culturing requirements, transport times, and human error. PCR testing has also been proposed and proven as a viable alternative, but culturing remains the recommended method for medical and reference laboratories [209–211]. Due in part to the much lower prevalence of the disease, the development of RDT devices for diphtheria have not become widespread.

Once diagnosed, diphtheria treatment has remained remarkably unchanged for decades, recommending a treatment of penicillin or erythromycin for two weeks, with a further course if required [209]. Other macrolides are available, including azithromycin or clarithromycin [209]. This must be given alongside the rapid administration of diphtheria antitoxin for toxigenic infections [209]

1.4.3 Vaccines and immune therapy

As impactful as the development of the diphtheria vaccine was, the development of the diphtheria antitoxin was rightly hailed as a major achievement at its time. Emil von Behring, with the aid of Shibasaburō Kitasato, discovered that the extracted serum of animals immunized with *C. diphtheriae* could be used to counteract the effects of diphtheria toxin in infected humans [212]. This work went on to win von Behring the very first Nobel Prize in Physiology or Medicine in 1901 [213]. Diphtheria antitoxin is still produced using equines today [214].

Despite the antitoxin's obvious success, it has become increasingly difficult to obtain in recent years, especially in high income countries. In Spain and Belgium, unfortunately, two highly similar situations presented themselves in 2015 and 2017 respectively [215,216]. In both cases, young children (6 and 3) who did not have the full protection of the vaccine, contracted diphtheria and were taken into hospital [215,216]. Tragically for both, delays in obtaining and administering anti-toxin treatment led to their deaths [215,216]. Reviews years before in 2009 and 2014 had warned that numerous countries across Europe did not carry stockpiles of the antitoxin due to the perceived low risk of disease occurrence, despite case imports occurring sporadically across the region [217,218]. Additionally, due to the low amounts being bought, many pharmaceutical companies across the globe have ceased production [217]. Due to the potential side effects of serum sickness, as well as the pressing lack of availability, alternative methods to diphtheria antitoxin have been proposed, including human mono- and polyclonal antibody-based therapies [214,219].

Developed during the 1920s and 30s before being fully introduced during the 1940s, the diphtheria vaccine is a toxoid formulation designed to offer immunity to diphtheria toxin (DT) [220]. At the time the vaccine was developed, case numbers were still high, with the USA alone recording 100,000-200,000 cases, and 13,000 – 15,000 deaths per year [221]. Cases fell drastically after widespread vaccine deployment, down to 4,333 in 2006, the lowest number since the WHO started recording [222]

The toxoid formulation has been shown to not be fully pure, with proteomic analysis suggesting additional immune benefits from the *C. diphtheriae*-derived cellular impurities that remain during production [220]. The vaccine is part of most national immunisation programs, often in the form of trivalent formulation with pertussis and tetanus. In the UK, a hexavalent 6-in-1 vaccine is given at 8, 12 and 16 weeks, incorporating vaccine components for diphtheria, hepatitis B, H influenzae type b, polio, tetanus, and pertussis [223,224]. It has been noted though that in many areas, especially Low-to-Middle Income Countries (LMICs), national immunisation programs have been heavily impacted by COVID-19, raising the spectre of large-scale vaccine gaps in regions most at risk of outbreaks.

1.4.4 Epidemiology

The WHO records reported diphtheria cases from multiple member countries, and guidelines suggest that nations monitor diphtheria cases at the district level [222,225]. While this is a simple task for many high-income countries, it can present a challenge for LMICs, especially in rural areas where cases might never make it to national records. In Europe surveillance is undertaken through The European Surveillance System (TESSy) managed by the European Centre for Disease Prevention and Control [226]. Previously there were multiple versions of a

diphtheria specific system, including the European Laboratory Working Group on Diphtheria (ELWGD) (1993 – 2006) and DIPNET (2006 – 2010), before being renamed the European Diphtheria Surveillance Network and incorporated into TESSy from 2010 onwards [226]. In the USA, the National Notifiable Diseases Surveillance System (NNDSS) encompasses ~120 diseases, diphtheria among them [227]. There is a risk that large scale networks could miss small but important changes in cases of diphtheria disease not considered a danger or priority, and one that must be accounted for when analysing data.

Post-vaccine introduction, diphtheria cases around the world reported to the WHO fell significantly [222]. Figure 1.11 shows the WHO reported case numbers per year.

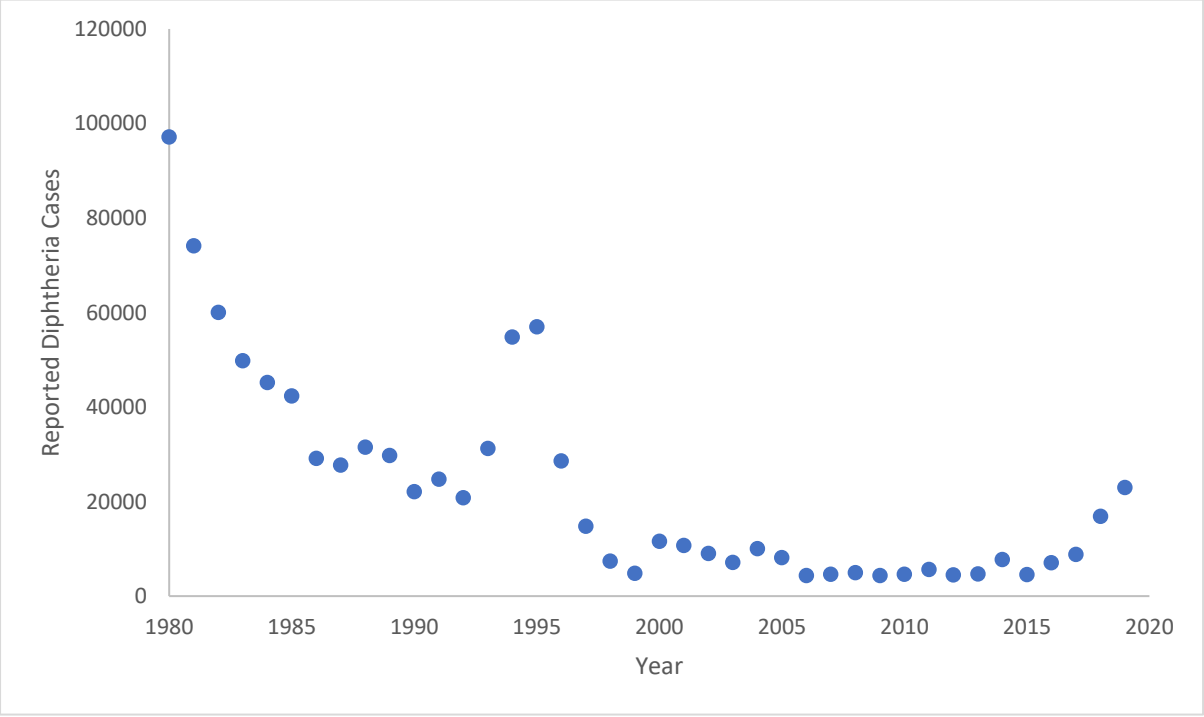

Figure 1.11: *Diphtheria case numbers reported to the World Health Organisation, 1980 – 2019* [222].

In the mid-1990s, there was a large spike in cases driven by the dissolution of the Soviet Union [130]. Gaps in healthcare infrastructure and delays in vaccination schedules created a perfect storm for diphtheria outbreaks in unprotected individuals, with Belarus becoming the case study example of outbreaks going forward [130,228]. Cases began in 1992, and doubled each year until 1995 when the outbreak peaked across the Post-Soviet states [130,222,228]. The epidemic was curtailed by numerous mass vaccination programs for all adults and children across the country, and only 36 cases were presented in Belarus by 1998 [228].

Following this, low incidence rates continued throughout the early 2000s and early 2010s, with single country outbreaks causing minor spikes [222]. However, over the past few years, as shown in Figure 1.12, case numbers have been on the rise once again [222].

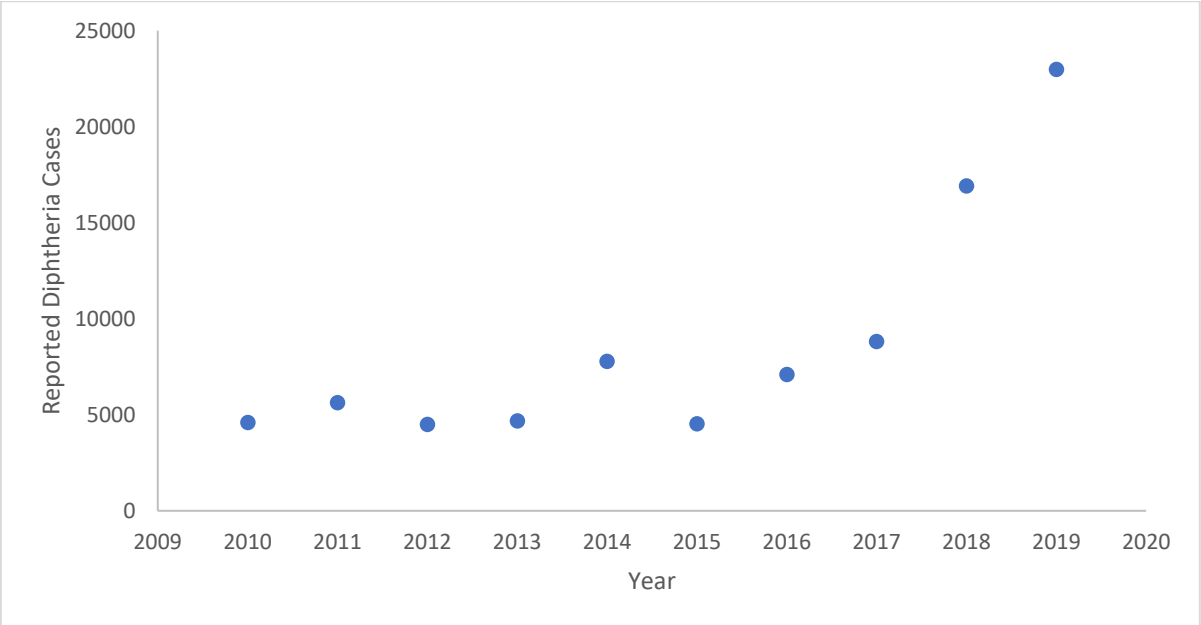

Figure 1.12: *Diphtheria case numbers reported to the WHO, 1910 – 2019* [222].

This significant rise in cases over the past 10 years has been driven by major outbreaks, in countries including Ethiopia, Madagascar, and Yemen [222]. In 2018, 16,911 cases was already the highest number since 1996 (28,624) [222]. As shown, 2019 continues this trend, increasing by over a third to 22,986 reported cases [222]. Figure 1.13 shows the 10 countries with the highest number of cases from 2010 – 2019, while Figure 1.14 shows a heatmap of cases by country for 2019.

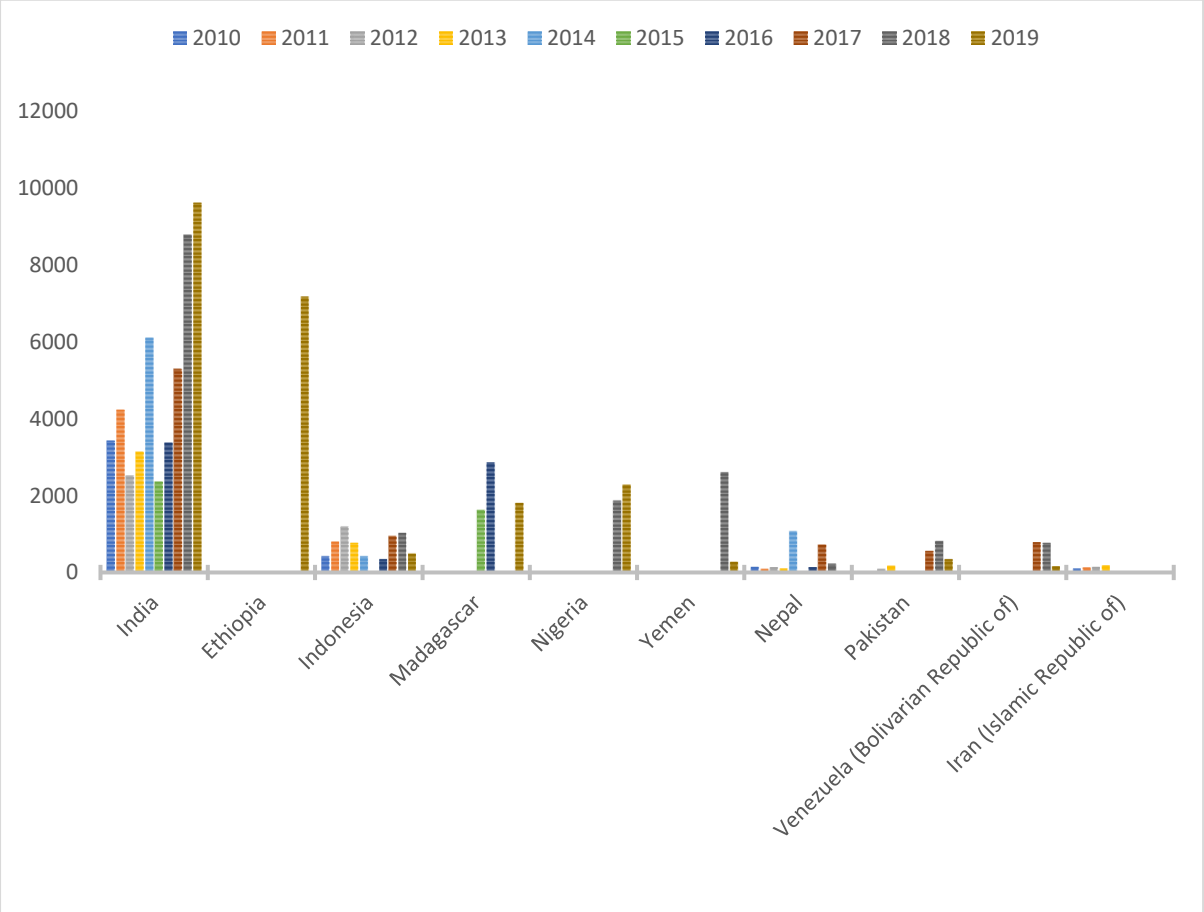

Figure 1.13: *Countries with the highest number of diphtheria case numbers reported to the WHO 2010 – 2019* [222].

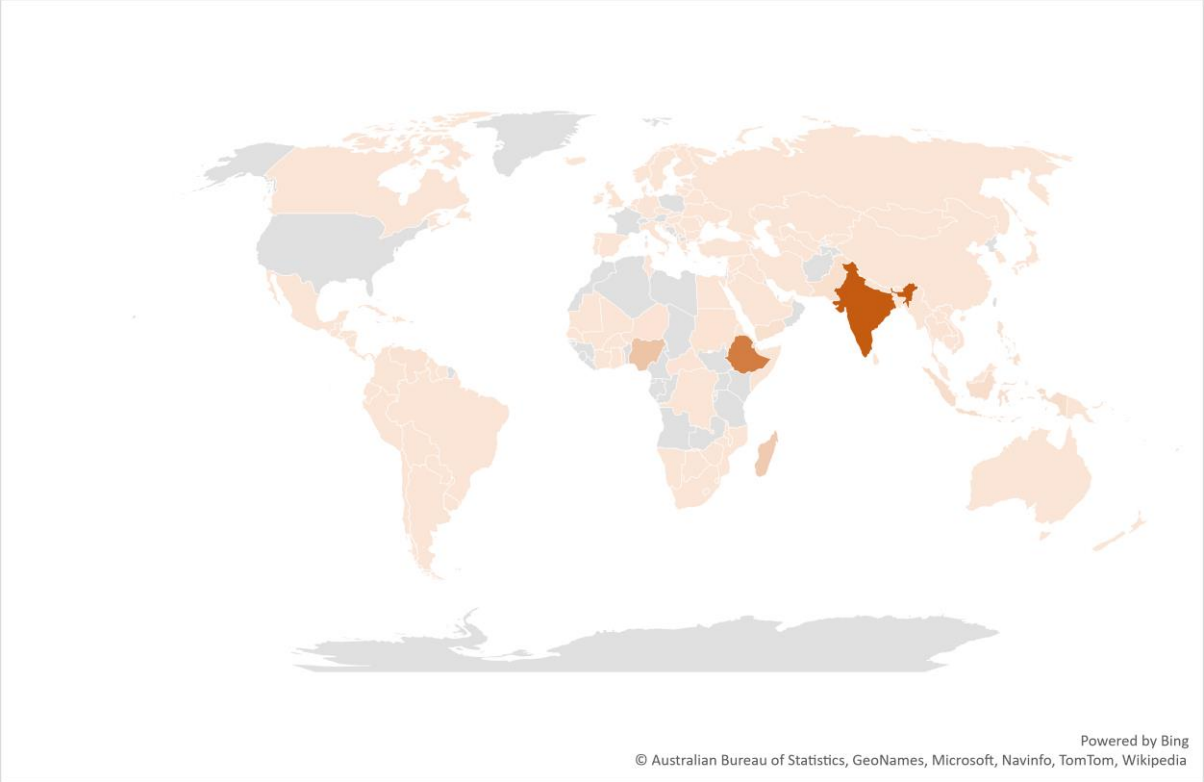

Figure 1.14: *Heatmap of diphtheria case numbers reported to the WHO in 2019 (darker colours show a higher number of cases)* [222].

Since the dawn of the 21st Century, India has reported an average of 61% of world cases per year, remaining the country with the largest levels of diphtheria [222]. This average did not change over the last ten years, with the 2010 – 2019 average remaining 61%. During this period, case numbers decreased initially before increasing again over the last few years. 9622 cases (42% of the global total) reported in 2019 is the highest number since 1991 [222]. In some high-income countries it is a similar story with reports in recent years being unexpectedly high. The UK reported 15 cases in 2018 and 12 in 2019, and while much lower than many other countries, they still represent the highest case number on record [222]. Germany cases were 26 in 2018 and 15 in 2019, the highest since 1982 [222]. While France reporting record is sporadic, the 39 cases reported in 2013 – 2017 (the most recent records) are almost as high as

the 45 reported in all years previously [222]. While this trend is not present in all areas of the globe, it is still an alarming statistic, and one that highlights diphtheria as a re-emerging global health threat.

1.4.5 Impact of genomics on our understanding of diphtheria

Compared to cholera, diphtheria has not had many 'landmark' papers in the field of genomics. Indeed, the two largest studies to date are Grosse-Kock *et al.*'s 2017 work on the Belarusian outbreak isolates, and Hennart *et al.*'s 2020 work on historical French and French territory isolates [130,138]. While our understanding of *C. diphtheriae*'s physiology and laboratory behaviour is relatively well developed, large scale analysis of genomes using modern genomic techniques is still lacking. Most current studies have focussed on individual outbreaks, and the combination of laboratory techniques and genomics, when used, can offer information on outbreak structure, virulence, and AMR. All these are important for outbreak management, but additional information such as the likely source, method of introduction, or how the outbreak isolates fit within a wider context is often unavailable due to the limited use of reference and representative genomes. This point to a large gap in information that could be highly relevant for future surveillance and control of diphtheria.

Even within *C. diphtheriae* genomics, there are differences in points of view between groups in key areas, such as the value of biovars, or whether certain isolates are even *C. diphtheriae* at all [198,199]. The creation of a shared framework in the vein of Mutreja *et al* for cholera is rapidly needed, both to develop and present a population structure that allows future groups to put their isolates in the correct context, but also to define key methodologies and terminology. This will allow genomic studies of the future to be directly comparable in a way that many are not today.

1.5 Bacteriophages and their role in Cholera and Diphtheria toxins

As viruses like the common cold infect humans, a bacteriophage is a virus that infects bacteria. Bacteriophages predate the bacterial cells by injecting their own genome and aiming to command the bacteria's DNA or RNA replication structures to replicate themselves [229]. This process causes major problems for the bacterium, often leading to bacterial cell death [229]. Indeed, it is fitting to call bacteriophages the natural predators of bacteria, parasitising the bacterium, often leading to death, for their own replication and survival [230,231] Figure 1.15 shows the replication cycle of bacteriophages.

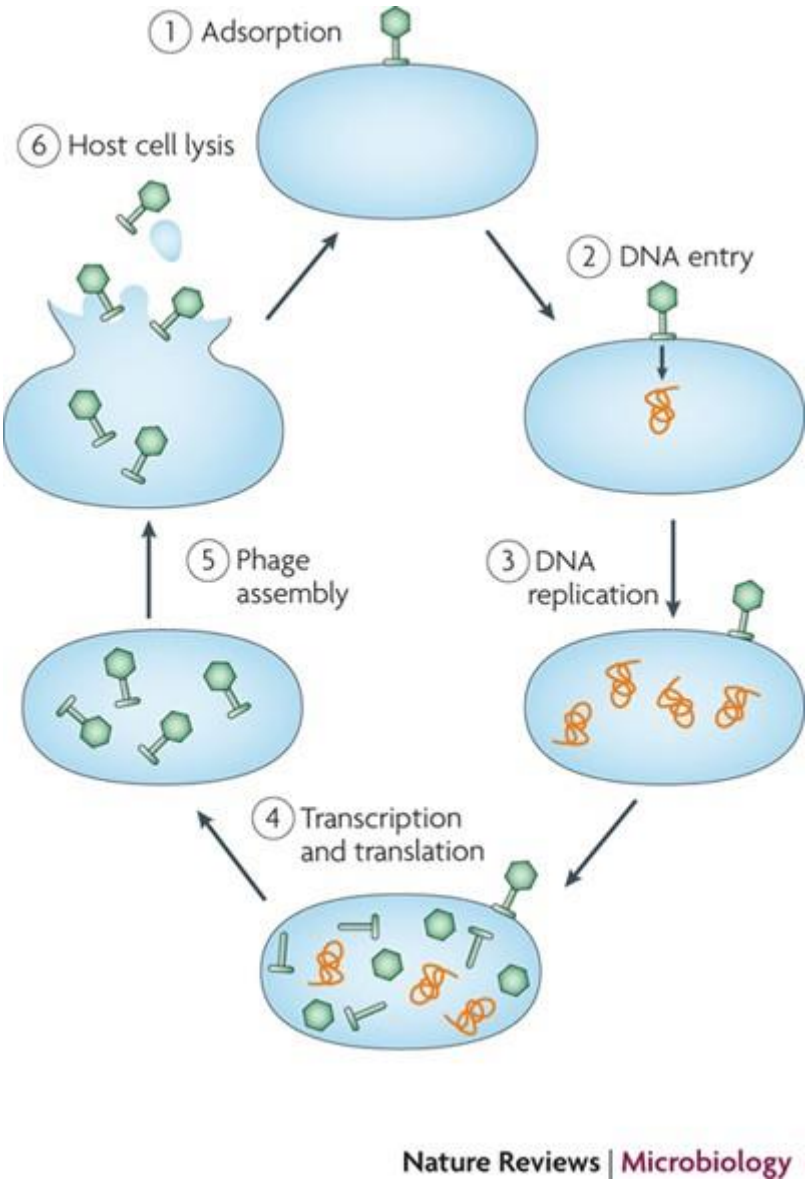

Figure 1.15: *The replication cycle of a bacteriophage, taken from Labrie et al, 2010* [229].

This replication is predominantly at no benefit (and usually great cost) to the bacteria, and the arms race between bacteriophages and their bacterial victims is forever progressing [229]. Absorption blocking, superinfection exclusion systems and the well-documented CRISPR–Cas systems are all utilised by bacteria to outright prevent infection, or to deal with the viral genome now-present in their genome [229]. One method demonstrated by *V. cholerae* is to secrete outer membrane vesicles that offer protection against predation [232]. Phages meanwhile

have adapted to counter many of these steps, and their continued success is testament to the rapid rate of their evolution across the order and families [229,233].

Phages are extremely diverse, which in turn is evidence of the variety of host targets available to them [229,233]. It is estimated that phages are the most numerous microorganisms on Earth, with 10 bacteriophages present for every bacterium [229,234]. The structure of phages has always been striking, and Figure 1.16 shows some of the major milestones in the imaging of these. Over 5000 phages have been imaged since 1959, and the high interest across disciplines was demonstrated when ϕX174 became the first sequenced genome in 1977, thanks to Fred Sanger and his team [21].

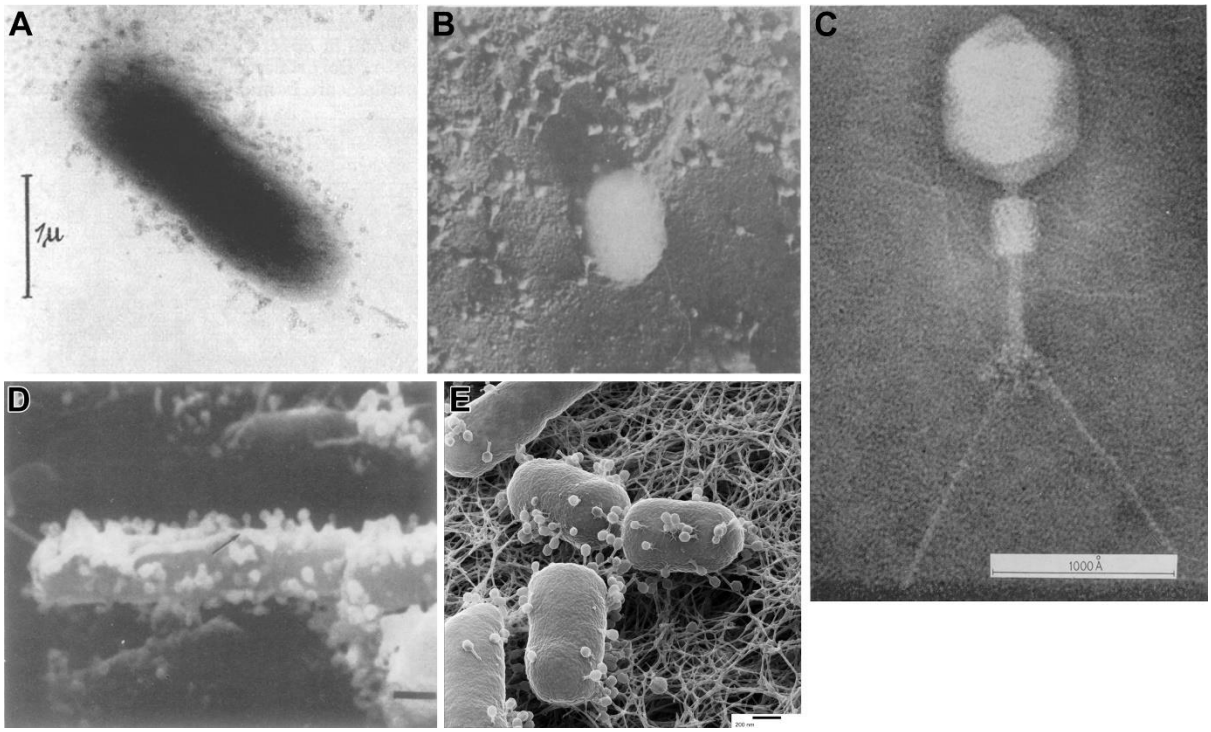

Figure 1.16: *Milestones of bacteriophage imaging, from 1940 – 2017. Taken from Almeida et al., 2018* [235]

Phages have been presented as a future antibacterial therapy, in an era of increasingly antimicrobial resistant pathogens [231]. By utilising phages as a carrier, antibiotic agents could be applied directly to bacterial cells, and could become part of the future of therapeutics [231]. The main barrier is not only future research, but also involves public opinion, policy and regulation in parts of the world [231].

Phage DNA can become integrated into the host cell genome, the bacteria then acquiring the genes contained within the virus [233,236]. This is the case in both *V. cholerae* and *C. diphtheriae*, where the toxin traditionally associated with both diseases is coded for by an incorporated bacteriophage gene [204,237]. This ability to secrete toxin presents an evolutionary advantage to both species, allowing both to weaken their host and its immune defences, albeit at the risk of host death. While the correlation between phage and toxin diversities are not fully understood, it is clear that while bacteriophages present a threat to bacterium survival, they can be utilised by the bacteria for their own benefit, whether by mechanism or accident.

Both cholera and diphtheria toxins are AB toxins, with DT one of the most well researched in the family [238]. The term 'AB' comes from the toxin family structure and mode of action, where an A-subunit and a B-subunit are synthesised and released by the cell bonded together by covalently associated disulphide bonds, also termed bridges [239,240]. AB toxins have a long history with bacterial infectious diseases, from anthrax and botulinum [241,242]. The latter, despite being one of the most toxic biological substances available, is regularly used in cosmetic surgery [243]. Cholera toxin belongs to a subfamily known as AB_5 toxins, where rather than one A and one B subunit, the toxin structure contains one A and five B, linked together non-covalently to the A subunit [240,244]. Other AB_5 toxins include those released by *Bordetella*

pertussis and *Shigella dysenteriae*. Outside of bacteria, AB toxins can be found in the kingdom *Plantae*, the best described of which is ricin, produced by *Ricinus communis*, the castor bean plant [245].

After synthesis and release by a bacterial cell, the toxin molecule enters the surrounding environment. Upon reaching a host cell, the binding subunit (B) binds with target receptors and allows the toxin to cross the lipid bilayer into the cell. Once there, the disulphide bond holding the subunits together are broken, and the now activated A subunit enters the cytoplasm, while the B subunit remains. Figure 1.17 shows a simplified graphical description of the process, from release to activation within a cell.

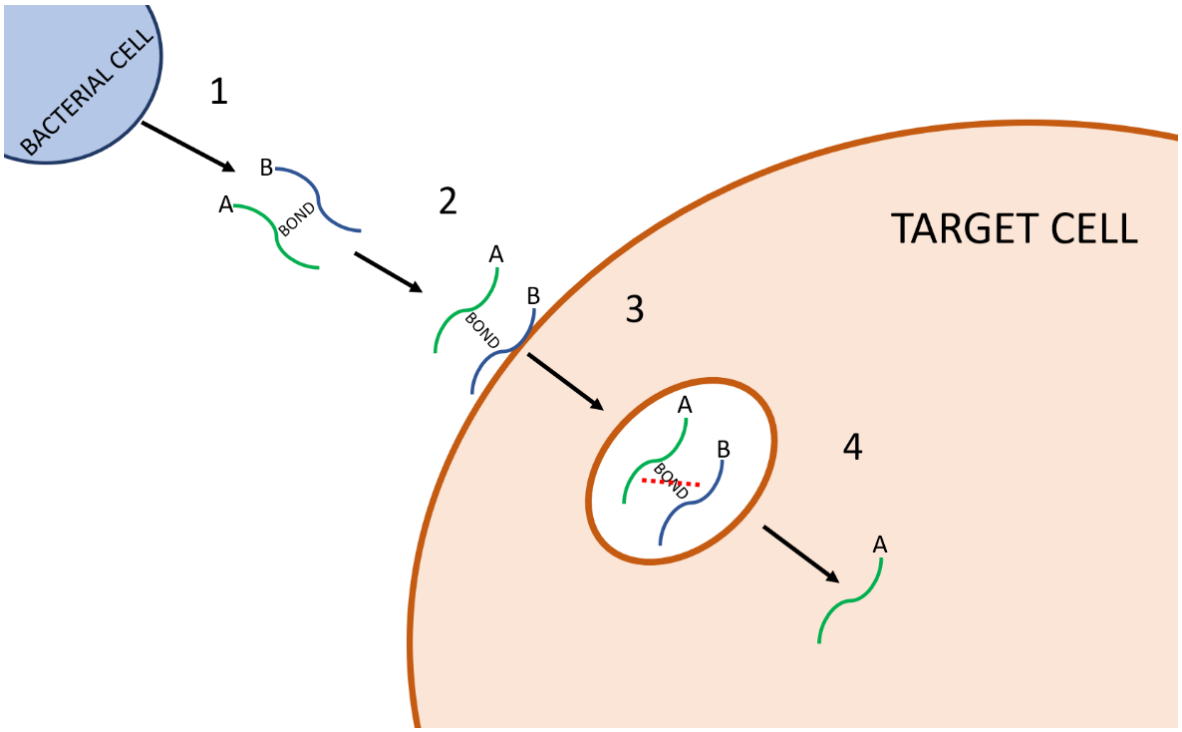

Figure 1.17: *Simplified graphical representation of an AB toxin, from release by the bacterial cell to activation within a host cell. 1) the AB toxin is secreted by the bacterial cell. 2) the B subunit binds to a receptor on a target cell's membrane and*

facilitates crossing. 3) now inside the target cell, the bond between the A and B subunits is broken. 4) The unbound A subunit is now an active toxin inside the cytoplasm.

1.6 Antimicrobial Resistance

1.6.1 History

Antimicrobial Resistance (AMR) is when a microbe (bacteria, virus, parasite, or fungi) develops the ability to no longer be susceptible to antimicrobial agents. The phenomenon was predicted by Alexander Fleming in 1945 [246,247]. Speaking to the New York Times, Fleming spoke about his Nobel Prize-winning discovery of penicillin, and how it could lead to selection pressures on bacteria if it was misused [246,247]. Unfortunately, he was completely correct, as numerous microbes have demonstrated very clearly in recent times. As the use of antibiotics has skyrocketed in the eight decades since Fleming gave that interview, the pressure to develop mechanisms of resistance has been ever-increasing for bacteria [248]. The constant challenge for any new antibiotic agent is that the pressure to develop or acquire resistance against it begins the moment it is first deployed [248].

The first recorded cases of resistance were reported in the late 1930s, shortly after the introduction of sulphonamides in 1937 [248]. Before penicillin was even deployed therapeutically, Abraham and Chain had reported a bacterial enzyme that broke down the drug, rendering it ineffective [249]. Resistance only became more widespread after its use became more extensive, the trend going hand-in-hand [248]. Many new antibiotics were discovered and introduced over the years, but in recent decades few new antibiotic classes have been discovered, as shown in Figure 1.18. No new classes means that on paper the arms race between the development of new methods of resistance and the development of new antibiotic agents has been all but lost. There are however many new antibiotics in active

development and trials. In 2019 the CDC reported 42 antibiotics in development, with one in four a novel class or mechanism of action, highlighting the continue efforts of researchers across the world to continue the fight against resistant pathogens [250].

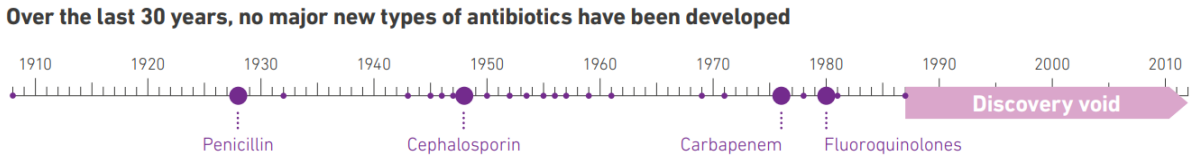

Figure 1.18: *Antibiotic discovery timeline, adapted from the WHO Antimicrobial Resistance Global Report on Surveillance 2014 infographic* [251].

In WHO first report on the global status of antibiotic resistance, published in 2014, species including *K. pneumoniae*, *S. aureus*, *N. gonorrhoeae* and *M. tuberculosis* were specifically mentioned as of serious concern [252]. Additionally, the report mentioned malaria, HIV and influenza as non-bacterial diseases that demonstrate serious levels of resistance [252]. All of the bacterial species mentioned had demonstrated widespread resistance by 2014, including untreatable gonorrhoea being reported in 2012 [252,253].

The threat posed by AMR has been acknowledged at national and international levels. The WHO ranks AMR as one of the biggest threats to global health, food security, and development today [254]. The European Commission adopted a One Health action plan against AMR, at the request of EU countries [255]. In the UK, the O'Neill report published in 2016 detailed the threat of AMR globally, commission by the then-UK Prime Minister [256]. As Figure 1.19 shows, O'Neill predicted that by 2050 there would be 10 million deaths

attributable to AMR [256]. In 2019, the UK Government published a five year national action plan to tackle AMR, focussing on the key areas of reducing and optimising antibiotic use, alongside innovation, supply and access investment [257]. Funding pots like The Fleming Fund have been set up, directing £265 million across 24 countries to further research into AMR globally [258].

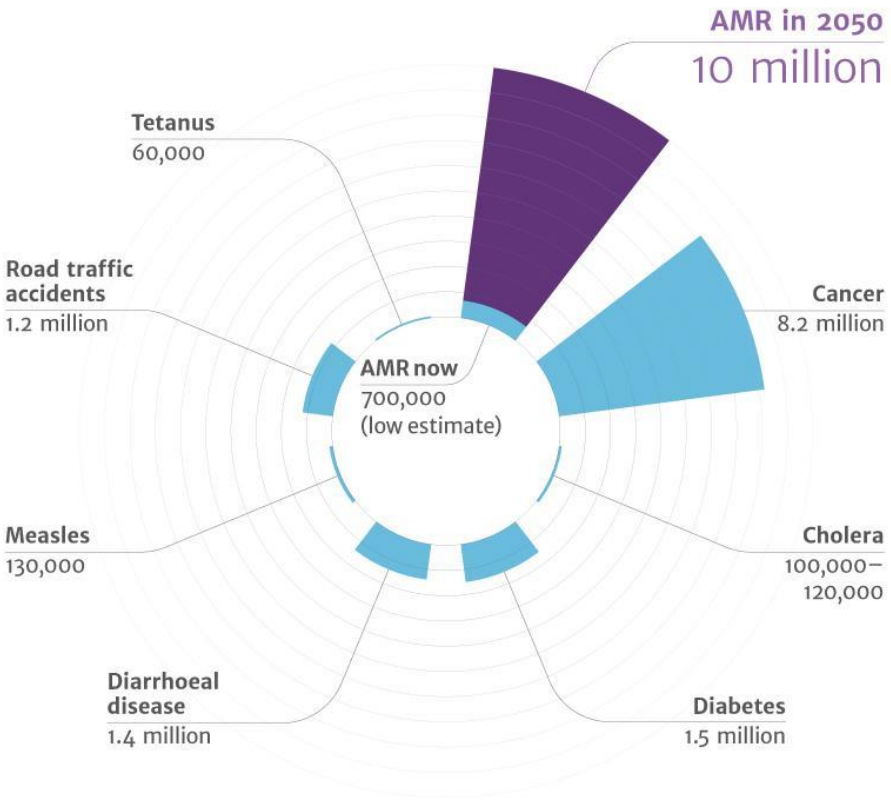

Figure 1.19: *Deaths due to AMR and other major causes, taken from O'Neill, 2016* [256].

1.6.2 Methods of spread

The most well-known way AMR is transmitted is via plasmids. The term 'plasmid' was first coined by Joshua Lederberg in 1952, and was an attempt to create a catch-all term for the

multitude of words being used at the time to describe ‘extra-chromosomal hereditary determinants’ [259]. His attempt was very successful, as proven by the widespread use of the word today. Another part of his definition was that ‘the plasmid itself may be genetically simple or complex’, and this was another element he got correct; while some plasmids are small, others can be much larger, both in size and gene content [260,261].

William Hayes was the first to show gene transfer without reciprocal genes going back the other way during exchanges between bacteria [262]. Initially published in Nature in January 1952, it was not until during a presentation in Pallanza, Italy at the Second International Symposium on Microbial Genetics that the so-called ‘Pallanza Bombshell’ discovery was widely received [262,263]. The mechanism of horizontal transfer is the main method of AMR spread that can act as a major driver of evolutionary change in a bacterial species [264]. Theoretically, all types of genes can be transferred from one bacterium to another, although certain types of genes, and indeed certain species of bacteria, have a higher propensity to undertake or be part of horizontal gene transfer [265].

There are three main methods used by bacteria to horizontally transfer genetic material [266]. The first is called transduction, utilised by bacteriophages to multiply within a host cell before infecting additional bacteria [266]. These phages can carry AMR genes from a previous host and transport them to new prey in this way, providing a small benefit to the bacteria they eventually infect, assuming it survives [267].

Transformation is the second, and is the release and uptake of free DNA into and from the environment [266,268]. DNA can enter the environment in various different ways, released by cell disruption and cell death, although some species have a higher propensity to 'kick out' genetic material during standard life (these include species such as *Streptococcus pneumoniae*, *B. subtilis*, *Acinetobacter calcoaceticus*, and *Pseudomonas aeruginosa*) [266,269–271]. DNA is not immediately degraded outside of the cytoplasm, and has been shown to last for hours or even days depending on the outside conditions [266]. Transformation is a mechanistically challenging for bacteria, requiring the cell to be in a state of 'competence' – able to uptake and integrate foreign DNA into their chromosome. To this end, they utilise a network of proteins and transport systems to transport the DNA through their wall and into their cell itself [268]. Some species are highly competent and able to undergo transformation often, while others are only competent under specific conditions, if at all [268].

Conjugation is the third method, and the one utilised most often in the movement of plasmids, and thereby AMR [265]. As shown in Figure 1.20, conjugation occurs when inter-cellular bridges are formed, and DNA passes across it [265].

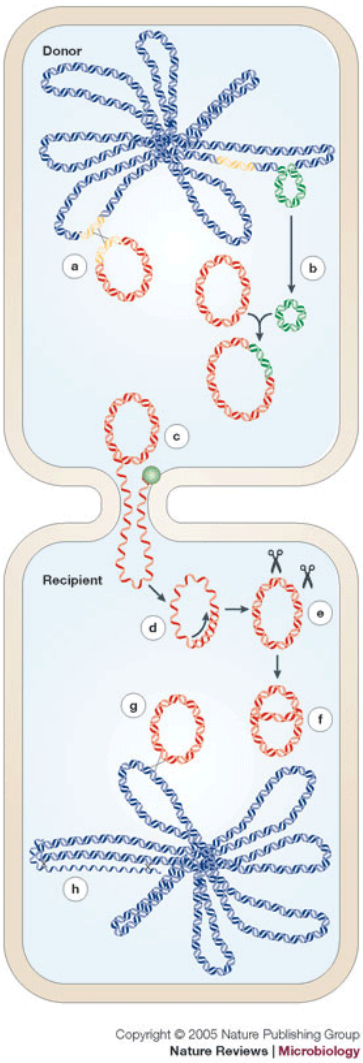

Figure 1.20: *Conjugation of plasmid during horizontal gene transfer between two bacterial cells. Taken from Gogarten and Townsend, 2005* [265].

Transmission of plasmids happens quickly at ambient temperatures, keeping the plasmid whole and unfragmented [265]. While some plasmids are genus specific, interspecies horizontal gene transfer allows AMR to cross species barriers regularly [267]. Plasmids have been found across the spectrum of bacterial hosts, including in humans and animal-associated microbes as well as those dwelling in soil and marine environments [267]. The ever-increasing use of antibiotics has driven the selective pressure to develop resistance to an all-time high, and

AMR is becoming so well transmitted around populations that it has serious implications for future treatment [272].

1.6.3 Impact on future treatment

The increasing levels of AMR being reported, coupled with resistance development to an increasing number of those drugs kept as 'last resort' options, raises a near-future situation where infections treatable today become entirely untreatable tomorrow. As shown in Figure 1.18, the O'Neill report predicted 10 million deaths due to AMR by 2050, and the UK's Special Envoy on Antimicrobial Resistance Dame Sally Davies (previously Chief Medical Officer, 2011 – 2019) recommended that AMR be listed alongside pandemic influenza and terrorism on the UK government list of threats to national security in 2013, highlighting the scale of the problem faced [273,274].

The untreatable future has already arrived in some cases, as previously mentioned completely untreatable gonorrhoea was reported as early as 2012 [252,253]. In *S.* Typhi multi-drug resistance appeared in 1972 across Mexico [275]. Presenting resistance to chloramphenicol, tetracycline, streptomycin, and sulphonamides, these gene came courtesy of a plasmid [275]. A similar plasmid family was found to offer MDR to *S.* Typhi decades later in Vietnam [275]. Increasing AMR in *S.* Typhi has led to the emergence of extensively drug resistant lineages in Pakistan, limiting treatment to only a handful of antibiotic agents. Azithromycin being the only broadly efficacious oral antibiotic treatment, while meropenem and tigecycline are the only other options, both requiring intravenous administration [275–278]. Unfortunately, Carey *et al* reported

azithromycin resistance in Northern India March 2021, as well as highlighting azithromycin-resistant *S.* Typhi bacteria in Bangladesh, Pakistan, and Nepal [278].

The picture presented by *S.* Typhi and *N. gonorrhoeae* is one that is being repeated by other pathogens across the world, and this will require new approaches and different ways of treatment going forward. Synthetic antibiotics (those engineered in a laboratory, rather than based on a natural project) have been proposed as one avenue forward [279,280]. Synthetic antibiotics can include existing structures adapted into new forms, or completely novel structures that do not exist naturally [279].

While the benefits of new antibiotics are easily understandable, there have been questions raised around the environmental impacts [280]. The high levels of antibiotics found in soils and marine environments show that antibiotic 'runoff' is an ever present danger, and one that drastically increases the development of AMR [280]. What impact these novel synthetic antibiotics might have on the ecosystems in these environments has been raised as an area needing further investigation [280].

Phage therapy meanwhile is an ever-increasing body of research that aims to harness bacteriophages and direct them to kill pathogens [281–283]. Mechanisms of action include using phages as a delivery system transporting antibiotic agents to specifically targeted species, utilising phage-derived enzymes to kill bacteria, and utilising bacterial phage-resistance mechanisms as potential new avenues to treat them [281]. As phage therapy has expanded, it has been met with policy challenges [231]. While some countries have begun using phage therapy

within their healthcare systems, they remain the exception [231]. What is clear is that the future of bacterial pathogen treatment must move beyond relying on antibiotics, exploring, and improving alternative methods while there is still time before they are all that is left.

1.7 Genome Sequence Impact on Policy

1.7.1 Cholera & diphtheria

Genome sequencing has expanded our understanding of many diseases, with cholera being a very good example. As whole genome sequencing has become the norm in both large- and small-scale investigations of *V. cholerae*, our understanding of the global population and its' spread continues to improve [123,124,192]. The shared framework and terminology have allowed transmission networks to be determined, which have a very clear implication for policy makers in single countries as well as across regions [123,124,192]. Inside a single country, knowing the source of an outbreak, or indeed if there are multiple outbreaks occurring simultaneously, facilitates a more targeted, rapid response, including cutting transmission networks and sending resources to affected regions. Additionally, the importance of cholera vaccinology has been highlighted and further developments may be aided by the wealth of whole genome sequences available [182,284].

Building upon the wealth of genomic research into cholera, the Global Task Force on Cholera Control (GTFCC) was revitalised by the WHO in 2011 [285,286]. In 2017 the GTFCC made a declaration aiming to eradicate cholera, and committed to targeting a 90% reduction in deaths caused by the disease by 2030 [287,288]. In the 2017 declaration and control strategy, the GTFCC set out the roadmap to 2030, based around early detection of outbreaks and quick responses to contain them. A goal was to prevent cholera recurrence using multi-sectoral approaches, and utilise effective mechanisms to coordinate experts across disciplines with local and global partnerships and the mobilisation of resources [288]. Thanks to the work of

numerous groups across the world, genomics has improved our understanding of the origin of cholera transmission routes and evolution across the globe. This information has been invaluable in the fight to eradicate this preventable disease, including identifying the importance of cholera focal points including the Bay of Bengal and Lake Victoria [120,123–125,289].

Diphtheria has not yet had the widespread genome sequencing and genomic analysis that cholera had benefitted from, but there is still room for policy impact. Diphtheria remains one of the diseases included on nearly all countries vaccination schedules, and recent outbreaks across the world have highlighted the continued threat of its re-emergence [193,208]. The large-scale gaps in antitoxin availability have also lead to policy change recommendations [217,218]. Major outbreaks such as in Yemen have shown the importance of controlling diphtheria, and that cannot be achieved without international collaboration [290,291]. As our understanding of *C. diphtheriae* globally grows, especially through genomics, these recommendations can be built upon and strengthened, to have a real impact on policy in regard to diphtheria.

Social turmoil are the main sources of major outbreaks of both cholera and diphtheria, and the political struggles within countries often lead to these perfect storms, such as the dissolution of the Soviet Union or the recent war and turmoil in Yemen [130,237,290]. If these challenges cannot be dealt with at a country level, international communities must come together to tackling them. This will reduce or even prevent further outbreaks and transmission of these diseases, a benefit to all. Outbreaks offer a fertile ground for further evolutionary developments in pathogens. such as AMR occurring at a faster rate than it otherwise would.

1.7.2 Other examples

The application of genome sequencing to track the spread of AMR has been instrumental in improving our understanding, as well as highlighting the scale of the problem [252,253,275–278]. By using genomics as one of the core pillars for highlighting the danger AMR poses to the future of global health, many countries have now begun to take meaningful action to counter the development and spread [255,257,258]. The WHO ranking AMR as one of the biggest threats to global health, food security, and development today shows the scale of international awareness of the problem, and this is due in part to the successful application of genomics [254].

Some organisations, such as the PHG Foundation, have championed the power of genomics impacting policy [292]. In their 2015 report by Luheshli *et al*, they highlight several other ways whole genome sequencing has come to bear in the world of UK policy [293]. One such case study highlights the challenge of *Streptococcus pneumoniae* vaccinology, where there are over 90 different serotypes that vary significantly in the dangers posed to infected individuals [293]. Current vaccine programs target only 13 serotypes [293]. Consequently, vaccine escape can and have occurred following inter-serotype recombination. This has been countered using whole genome sequencing genomics, where large studies have been able to determine the evolutionary changes that have led to vaccine escape, and making recommendations as to improvements [294–297].

Genomics can also be used to quickly understand pathogen outbreaks. Both in the swine flu outbreak in 2009 and the Ebola outbreak of 2014, transmission of the diseases were able to be

tracked almost in real time, as well as placing a time scale to the emergence of the pathogen and determination of its origin [298,299]. Both highlighted the power of genomics to understand ongoing pandemics, something that continues to be built upon during the ongoing COVID-19 pandemic.

1.7.3 COVID-19

As the largest pandemic in modern history, COVID-19 has, and continues to have, major impacts on the daily lives of billions. First reported in December 2019 as a viral pneumonia outbreak in Wuhan, China caused by a novel coronavirus SARS-CoV-2, COVID-19 has gone on to cause over 114 million cases and 2.5 million deaths, as of the March 2021 [300].

This has naturally had a major impact on policy in countries around the world as governments attempt to control outbreaks within their own nations, and genome sequencing has played a large part in improving our understanding of the pandemic globally. It is imperative to understand the evolutionary changes occurring in COVID-19 in real time [301]. This information guides the development of diagnostics, treatments, and vaccines, providing knowledge of sources and transmission networks, as well as future treatment failures and vaccine escape potential [301].

Acknowledging this importance, in the UK the COVID-19 Genomics UK Consortium (COG-UK) was founded [302]. Begun as a concept on the 4th March on a phone call between experts and enthusiasts of genome sequencing, this was quickly followed by a meeting on the 11th, and a proposal for funding on the 15th [303]. Funding began on the 1st April 2020, and was set

up as a joint initiative between the Department of Health and Social Care, UK Research and Innovation, the Wellcome Sanger Institute, public health agencies and hospitals across the UK, and many academic partners [302].

Thanks to the early inception and tireless work of COG-UK, the UK leads the world in SARS-CoV-2 genome sequencing [304,305]. COG-UK passed 100,000 genomes within 8 months of the pandemic reaching the UK (November 2020) and has now sequenced over 350,000 genomes as of March 2021 [306,307]. The US follows in second for the number of genomes sequenced, while Australia and Iceland have the highest 'sequenced genome per case' ratio [305].

This ever-increasing amount of genomic data has allowed lineages of SARS-CoV-2 to be found, including some reported to have an increased rate of transmission [308]. The large amount of data has allowed researched to establish lineage dynamics across the UK, with over 1,000 lineages having become established in the country, as reported by Du Plessis *et al*. in February 2021 [309]. It is of paramount importance to understand the national and global picture of this disease, as this can then feed into policy makers and their decision making. These decisions can have massive impacts on people's lives, as choices of when to impose travel bans or lockdowns can save many from infection and potentially death, although they have massive impacts on people's livelihoods and mental health [310].

Most quantum leaps in global health have been driven by major public health crises [311]. Outbreaks of viruses like Ebola and Zika had primed the world to the threat of potentially

global serious disease pandemics, and coronaviruses had previously been raised as a potential source of concern, with Chen *et al.* raising the idea as early as 2007 [293,311,312].

1.8 Aims of the thesis

Utilising novel genome sequences from bacteria collected in India, Ghana, and across the globe, this thesis aims to demonstrate the power of genomics to understand pathogen evolution across time and space. *C. diphtheriae* and *V. cholerae*, despite their many obvious differences, both share a large number of important characteristics. Both Cholera and diphtheria have been described as ‘ diseases of poverty’, with low-income countries significantly more affected by the disease that those with higher incomes [313]. This is true in terms of total case numbers, rate of cases/100,000 people, and mortality, where the numbers in lower-income countries are many times higher than in higher-income countries [313].

Additionally, outbreaks of the diseases caused by these two bacterial pathogens are both strongly linked to the same socio-economic factors. Diphtheria, much like cholera, will re-emerge significantly in areas of social turmoil and among populations without access to adequate medical care. This has been recently demonstrated among the Rohingya refugees as well as the war in Yemen, where the forced displacement of millions and confinement in close proximity quickly caused a significant spread of cases – both for cholera and diphtheria [183,290,291,314–317]. In both situations, widespread vaccination efforts and the expansion of medical treatments were able to curtail and control the outbreaks, but both diseases will quickly return under similar conditions without correct management.

The power of genomics has been demonstrated in the field of *V. cholerae*, giving us a global framework from which multiple studies have grown. This has facilitated comparable country, region, and continent level cholera studies, something which research on diphtheria and *C. diphtheriae* has been largely lacking. Here, we aimed to use novel genomes from India, the country most affected by diphtheria, combined with previously published genomes, to elucidate a global picture of the *C. diphtheriae* population structure. Additionally, while AMR has not previously been reported as a widespread problem within the species, it is important to understand if the situation is changing. Additionally, due to the importance of diphtheria toxin in diphtheria pathogenesis, variation within the diphtheria toxin-encoding gene *tox* warrants investigation. An aim was to exploit novel Indian *C. diphtheriae* genomes within a national context to better understand the population structure, AMR, and toxin variation within the country most affected by *C. diphtheriae*.

A further aim was to investigate, using genomics, the evolution of cholera within a specific country, Ghana, and in the context of antigenic variation e.g., the evolution of the O139 serotype. Hence, we investigated the national picture of cholera in Ghana, with representatives from other African cases and beyond to place Ghana *V. cholerae* into the global picture of the species. We investigated the presence of AMR in the *V. cholerae* within the country and explored the impact outbreaks in neighbouring countries may play in relation to cholera in Ghana.

Finally, the emergence of O139 *V. cholerae* presents a fascinating story, going from being perceived as an incredibly competitive lineage purported to be the start of an eighth pandemic to obscurity in only a few decades. By assembling a collection of genomes from O139

isolates, we aimed to improve our understanding of how O139 was able to so effectively outcompete O1 cholera and determine what drove this seemingly very fit lineage to almost entirely disappear.

Across both species, these results could have implications for scientific, medical and policy stakeholders at national and international levels. Our aim was to reveal the population structure, AMR, and toxin variations across these two species than play incredibly important roles in global health today. Despite the number of cases, both pathogens still need to be controlled with the right resources, and these studies could add more information in the fight to eradicate these preventable diseases.

2.0 Methods

2.1 *Corynebacterium diphtheriae* and *Vibrio cholerae* genome data collection

To assemble our *C. diphtheriae* and *V. cholerae* collections, novel clinical isolates were grown and DNA was extracted, before short read Whole Genome Sequencing (WGS) was carried out using an Illumina HiSeq v4 platform at the Wellcome Sanger Institute. These were combined with publicly available genomes and their metadata obtained from the National Center for Biotechnology (NCBI) Genbank and European Bioinformatics Institute (EMBL-EBI) European Nucleotide Archive identified after reviewing previous studies [288,318].

2.2 *C. diphtheriae* core gene analysis

After the extremely high recombinogenic variation was identified across the *C. diphtheriae* genome collection (identified using Gubbins (v2.4.1) and Phandango (v0.9)), we used Prokka (v1.5) to annotate our genomes, before Roary (v3.13.0) identified the genes present within 99% of our genomes, extracting and concatenated them into a core gene alignment [319–322]. SNP-sites (v2.5.1) was used to identify the single nucleotide polymorphisms (SNPs) present across our core gene alignment, and this was the base of our phylogeny.

2.3 *V. cholerae* genome mapping

To create the alignments we would use in the phylogenetic tree construction for both the Ghanaian and O139 *V. cholerae* analyses in Chapters 5 and 6, we mapped all our genomes against the reference genome *V. cholerae* NI6961 utilising SMALT mapping and the 'multiple_mapping_to_bam' and 'join_dna_with_indels' scripts developed by Simon Harris [323,324]. The alignments produced by these scripts were filtered for recombination using Gubbins (v2.4.1), and the SNP alignment output from Gubbins was used for the construction of our Ghanaian and O139 phylogenies [321].

2.4 Phylogenetic tree construction

Across Chapters 3, 5, and 6, the methods of constructing phylogenetic trees were kept constant. IQ-TREE (multicore version v1.6.10) and the inbuilt ModelFinder were used to infer the phylogenetic structures of all three SNP alignments (*C. diphtheriae*, Ghanaian *V. cholerae* and O139) [102,325]. 1000 pseudo-bootstraps were used during the construction of our phylogenies.

The phylogenetic trees produced were annotated with geographical and temporal metadata, as well as results of further analyses, using Figtree for ordering of nodes and the Interactive Tree of Life (iTOL v5.5.1) for visualisation [112,326].

2.5 Genotypic AMR gene analyses

Our *C. diphtheriae* genome collection was interrogated for the presence of AMR gene determinants using ARIBA (v2.14.6) and the ARG-ANNOT AMR gene sequence database [104,327]. The O139 *V. cholerae* genome collection was analysed using the Short Read Sequence Typing for Bacterial Pathogens tool (SRST2 v0.2.0), while the Ghanaian *V. cholerae* collection was analysed with both ARIBA and SRST2 after Kraken2 (v2.0.8), KrakenTools, and Abacas (v1.3.1) were used to mask and remove reads that were not identified as *V. cholerae* [105,328–330].

2.6 Diphtheria toxin gene *tox* analyses

In-silico PCR, utilising a script written by Simon Harris, was used to extract the diphtheria toxin gene *tox* from our *C. diphtheriae* collection [106]. Non-synonymous mutations were identified by comparing novel variants against that carried by the vaccine strain and plotted onto the diphtheria toxin protein 3D model 1XDT obtained from the Protein Data Bank utilising the UCSF ChimeraX (v1.1) tool [331–333]. The level of impact that these mutations may have on the protein structure was estimated by PHYRE2 (v2.0) and the inbuilt SuSPect [334,335].

To assess if the variation within the *tox* gene was a result of diversity across the corynephage. We mapped 11 *tox*$^{+}$ complete *C. diphtheriae* genomes from NCBI Genbank to the corynephage sequence annotated in *C. diphtheriae* isolate NCTC 13129 using BWA (v0.7.17-r1188) [59,60]. This mapped alignment was used to create a maximum likelihood

phylogenetic tree using the same methodology stated in previous sections, using IQ-TREE and ModelFinder over 1,000 pseudo-bootstrap replicates, before annotation using iTOL.

2.7 Statistics

RStudio (v4.6.1) was used to carry out the statistical testing in Chapter 3 and 4. Pearson's product-moment correlation was used to determine the significance between the decade of isolation and the *tox* gene variety present within those genomes. Pearson's product-moment correlation was also used to determine if there was any significance between the average number of antimicrobial resistance genes per genome and their decade of isolation, and the number of non-toxigenic isolates per decade. A chi-squared test was used to assess the relationship between non-toxigenic and toxigenic isolates from HICs and LMICs [109,110].

2.8 Collaborations

In-vivo AMR analysis was carried out by Thandavarayan Ramamurthy and their team in India for the novel Indian *C. diphtheriae* isolates, and Japheth A. Opintan and their team in Ghana for the novel Ghanaian *V. cholerae* isolates. Karthick Vasudevan carried out the BEAST analysis on the *C. diphtheriae* collection in Chapter 3, and on the Ghanaian *V. cholerae* collection in Chapter 5, as well as the BAPS analyses [336,337]. Agila Pragasam carried out further *in-silico* gene analysis on the Ghanaian *V. cholerae* collection to determine if mutations were present within the Ghanaian *V. cholerae* isolates' *gyrA* and *parC* gene and which *ctxB* variants were carried by the isolates. The online tool CholeraeFinder was used for *gyrA* and *parC* analysis, followed by manually identification of the mutations sites and which

mutants they corresponded to [338]. *In-silico* PCR, utilising the same script written by Simon Harris as the *C. diphtheriae tox* gene analysis, was used to extract the *ctxB* cholera toxin genes present within our Ghanaian *V. cholerae* genome collection [106]. The *ctxB* results were compared to existing catalogues to determine which variant they were.

2.9 Data Files & Scripts

All data for genomes used throughout this thesis can be found here: Genomic Metadata Master File.

All scripts used for these analyses can be found here: Thesis Codebook

3.0 National and global genomic epidemiology of *Corynebacterium diphtheriae*

3.1 Introduction

There is an increasing recognition that a detailed understanding of the population structure of a microbial pathogen facilitates intervention programmes. A key component to gaining such an understanding is to define this structure using genomic analysis[66,69,70,120,123,124,339]. The causative agent of diphtheria, *C. diphtheriae*, has been predominantly analysed using classical laboratory techniques, and thanks to the work of scientists such as Klebs, Löffler, Roux, Yersin, Behring and Kitasato, the disease's symptoms and treatments have been defined for over almost 140 years [128,194–196,212].

Previous phylogenetic studies on *C. diphtheriae* and diphtheria based on genomics, have predominantly been outbreak-based, focussing on a relatively small number of samples analysed together with public genomes to act as reference points, although in some cases though no appropriate references were available [132,136,137]. This approach allows the population structure to be defined within an outbreak but the overall context can be lost. How outbreak isolates relate to previous outbreaks, both in the local region and beyond, can be incredibly important to public health officials and policy makers, and can potentially reveal avenues of transmission or potential sources of any outbreak. Additionally, it can be challenging to compare data and thus outbreaks, due to differing methodologies and terminology used in these different studies. This loss of data generally means that potentially crucial information on population structure is unavailable or not determined [132,136,137].

By collating publicly available genomes into a single large database as well as using a methodologies for analysis that are both open-access and relatively easy to replicate, this loss of information could be alleviated. An outcome could be a better understanding of the circulating *C. diphtheriae* population structure across the globe. As mentioned, this information is incredibly important for health care officials and policy makers and could lead to common terminology and methodologies, as has been created for other pathogens, such as *V. cholerae* [120].

The phylogenetic analysis of many smaller diphtheria outbreaks have been based on quite small numbers of genes, including those based on MLST [132,136,137]. For MLST, only six or seven housekeeping genes are chosen in part because the selected genes are highly likely to exist within the genome of a free-living bacteria, as well as their relatively conserved nature genetically. The sequences of these MLST prototype genes are then determined and compared to a database of previously defined allelic variants of these genes to determine an MLST designation. Rough phylogenies can be based upon these designations, or alternatively the MLST genes can be concatenated together to form small alignments. These methods are acceptable for a broad species such as *C. diphtheriae*, but the level of resolution is small compared to more comprehensive analysis based on larger gene sets. By using close to whole genome sequences based on next generation technologies, a much larger amount of genetic data can be compiled and then exploited for subsequent phylogenetic analysis, providing a much more accurate and detailed picture.

Using these whole genome sequences to create an accurate phylogeny of *C. diphtheriae* poses its own challenges. *C. diphtheriae* exhibits high levels of diversity, and this can make traditional mapping-based approach much less effective. Mapping to a reference genome remains the bedrock of phylogenetic alignments, but due to the diversity, the genetic distances of isolates from the nearest high-quality reference genomes is large in *C. diphtheriae* as a species. This can result in sections of the genome being lost for particular isolates, as they do not map effectively to the reference being used. This diversity is driven in a large way by recombination [127]. Tools are available to identify and remove recombinatorial regions, including the widely used tool Gubbins [321]. While this tool works well for highly similar and clonal species, it struggles with those that are highly diverse, and the requirement for a mapped alignment input means that there are many problems adapting this methodology to fit *C. diphtheriae*.

Time-scaled phylogenetic analysis, a common method used to estimate dates on key evolutionary events, also requires a mapped, recombination free multifasta alignment to facilitate accurate investigations. This is another challenge in the monitoring of *C. diphtheriae*, as the information tools like BEAST can provide can be lost or unattainable. Such information potentially includes the estimated dates for lineage branching events and transmission dynamics, all incredibly useful for disease monitoring and understanding outbreaks.

Due to these challenges, a phylogenetic analysis based on a maximised core gene approach is considered the most optimal for *C. diphtheriae*, and has been widely used is previous studies [130,138]. Sequenced genomes are optimised and annotated before the sequences of genes

present in the vast majority of them are extracted. These genes are then concatenated together into a new base sequence for each genome, before being combined together into one multifasta file and aligned. Recombination events cannot be removed from this type of alignment using currently available tools [321]. This is due to the sequence not being a 'real' genome, but rather many gene sequences stitched together for each isolate. These genes are now artificially placed next to each other, in a context that they would never be found in a real genome. This is a format recombination removal tools are not able to understand. As a hypothetical example, the tool may designate the second half of one gene and the first quarter of another as a section of recombination, when in reality those genes were hundreds of bases apart within the original genome. If such artificial sections of alignments are removed this way, it may potentially introduce errors into any final phylogeny.

AMR is an emerging challenge for the treatment of many infections. While AMR has not been considered a major problem in diphtheria to the extent of widely impacting treatment, it is nevertheless a potential threat, and there have been multiple reports in recent years of resistance being present in *C. diphtheriae* or in clinical settings where the disease is occurring. Using computational tools, the presence of genes within a genome that can confer resistance to antibiotics can be determined. Continuing to assess the presence or absence of AMR genes within a pathogen population, regardless of the perceived risk to current treatment, is now regarded as an imperative for disease monitoring. Novel events in AMR development, such as when a bacterium picks up new AMR genes that compromise current treatments, can theoretically occur at any time. Thus, it makes sense to know how widespread novel resistance might be before the point when clinical treatment starts.

We set out in this study with an aim to use whole genome sequences generated through the application of next generation technology to elucidate and understand the global population structure of *C. diphtheriae*. To this end, 502 *C. diphtheriae* genomes were collected from 16 countries and territories that represented 122 years of isolation, presenting a globally representative picture of the species. This collection included 61 novel genomes generated from isolates collected between 2015 – 2018 in India. Additionally, while the AMR burden across the species has not been previously considered to be a major problem, it is important to continue to investigate the AMR genes present within these genomes to be prepared for any future impacts on treatment. In addition to a global analysis, we also investigated the potential for temporal and geographical phylogenetic analysis on determined closely related groupings of *C. diphtheriae*.

Additionally, we aimed to gain a better understanding of diphtheria at the national level. By focussing on India, the country where over 50% of cases are reported, we placed the 61 novel Indian *C. diphtheriae* genomes within their national context, combining them with 61 previously sequenced isolates to create a 122 genome-strong collection. We also investigated the AMR burden of *C. diphtheriae* within the country. By using the same methodologies for the Indian collection as we had for the global collection, we were able to make direct comparisons between the resulting analyses, presenting conclusions about the molecular epidemiology of *C. diphtheriae* at both the national and international level.

3.2 Creating a global collection of *C. diphtheriae*

61 *C. diphtheriae* were isolated in India from clinically diagnosed cases of diphtheria over a four-year period covering 2015 – 2018. DNA from these isolates were sequenced at the Wellcome Sanger Institute using an Illumina HiSeq v4 platform producing short read data. A literature review of major outbreaks reports was also undertaken, and appropriate genomic sequences were extracted from the publicly available database National Center for Biotechnology (NCBI) Genbank [288].

441 genomes and their metadata were extracted from these publicly available sequences and they were combined with the novel Indian genomes to produce a collection of 502 *C. diphtheriae* genomes. With the oldest sample isolated in 1896 and the most recent in 2018, the genomes spanned 122 years of studies on diphtheria. Sixteen countries and territories were represented in the isolates, hailing from 6 continents. This collection provides an opportunity to examine the evolution and variation of *C. diphtheriae* across both space and time, and begin to better understand how the species' population structure is structured. Figure 3.1 and Figure 3.2 show the countries and years of isolation for all 502 genomes.

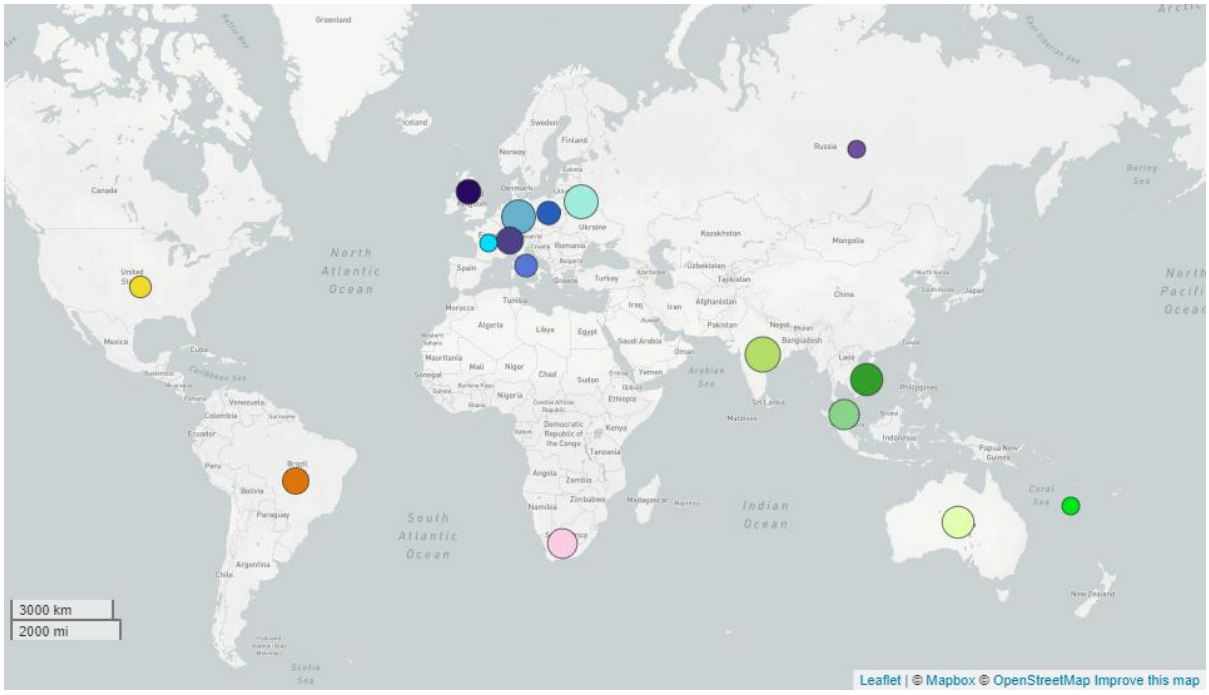

Figure 3.21: *Map of 502 C. diphtheriae genomes, coloured by country of origin and scaled by number of isolates.*

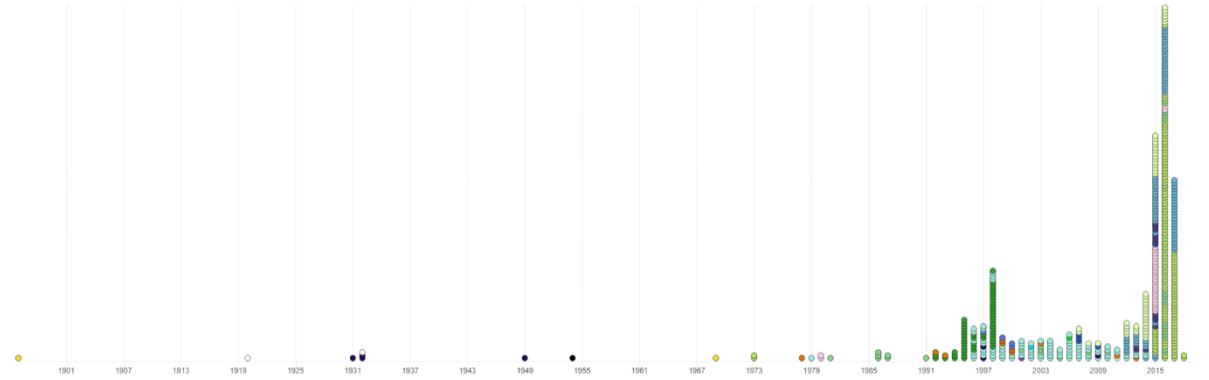

Figure 3.22: *Timeline of 502 C. diphtheriae genomes, coloured by country of origin.*

The majority of the *C. diphtheriae* in the collection were isolated within the last few decades, with 85 from the 1990s, 69 from the 2000s, and 319 isolated during the 2010s. Geographically, the majority of genomes originated from Europe and Asia, with India, Belarus, Germany, and Vietnam providing the highest number of individual isolates; 122, 84,

82, and 54, respectively. These are areas where diphtheria cases have been recorded with varying degrees of regularity. In Western European countries cases numbers remain low, but large-scale changes to conditions such as the dissolution of the Soviet Union and the subsequent impact on Belarus, have led to large increases in incidence. Outbreaks or individual cases have also been linked to the influx of people during the so-called 'migrant crisis' that affected many parts of Europe (e.g., Germany) in recent years. Other countries, including Brazil, South Africa and Australia have had sporadic flare ups of the disease, reporting such cases and sometimes sequencing genomes. Across Asia, countries such as India and Vietnam report diphtheria much more regularly, and case numbers are collectively relatively higher. In these countries, not all isolates are linked to genome sequences, often due to cost and the availability of other resources such as laboratory equipment or the lack of researchers to operate them.

By combining these differing data collected broadly across the globe into one collection, the population structure of *C. diphtheriae* can begin to be elucidated in more detail, and the recent evolution of the species can become clearer.

3.3 The global phylogeny of *C. diphtheriae*

Once the 502-genome global collection had been assembled, construction of a *C. diphtheriae* phylogeny could begin. As covered in the 1.2.5 section, the traditional method is to produce a mapped multi-genome alignment that can be used as the base of a phylogeny. However, this has not proven effective in previous *C. diphtheriae* studies. Instead, these previous publications have used core gene approaches, including core gene alignments and core

genome MLST, where new MLST genes are chosen from only those present within the core genome of the collection [130,132,137,138]. The main obstacle to mapping approaches is the high diversity of *C. diphtheriae*, a big part of which is the high levels of recombination reported across the species.

To determine the effectiveness of traditional mapping-based approaches coupled with recombination removal tools, we mapped our 502 genomic paired fastq files against the *C. diphtheriae* medical reference genome NCTC 11397. The multifasta alignment produced was then used as the input for Gubbins, to evaluate the effectiveness of this widely used recombination removal tool for *C. diphtheriae*.

Figure 3.3 shows the recombination pattern created using Gubbins and plotted using Phandango for the core gene alignment phylogeny. As shown, significant portions of the genomes were linked with recombination, due to the high levels of diversity as well as inefficient mapping.

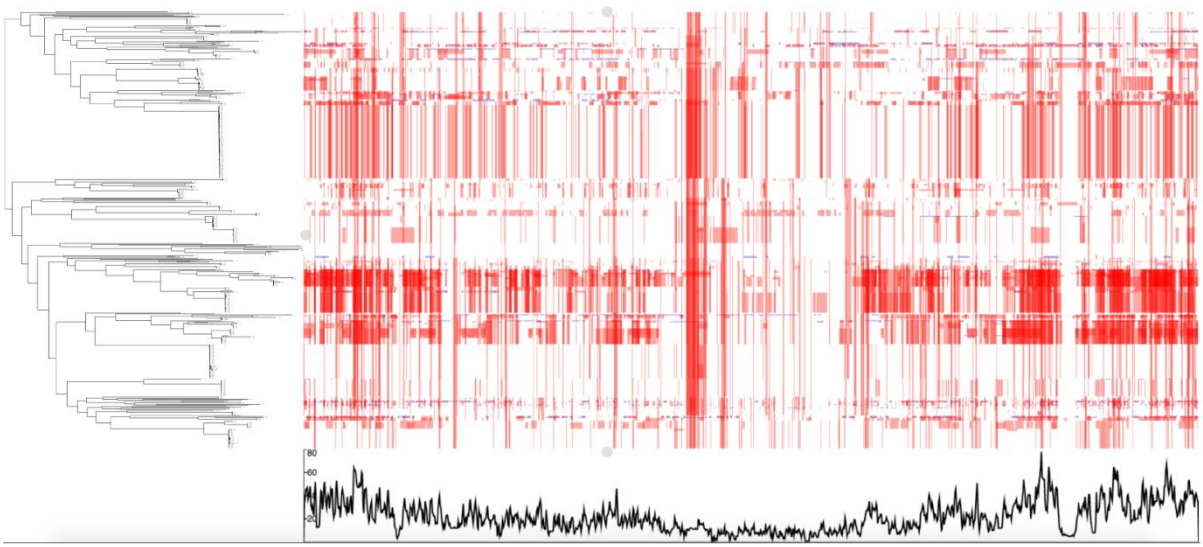

Figure 23.3: *Recombination patterns detected by comparing the reduced genomes of C. diphtheriae isolates, identified by Gubbins and plotted using Phandango* [321,322]*. Red colour blocks indicate ancestral events, while blue blocks indicate events only affecting a single leaf.*

These data emphasise the problem with using mapping-based approaches when constructing phylogenies of *C. diphtheriae* and highlighted the need to use alternative methods. Even though the methods chosen were somewhat able to deal with the high levels of recombination present across the collection, the approach would have impacted the final phylogeny and skewed any conclusions that might be drawn from it.

The core gene approach widely used in previous studies was chosen as an effective alternative. By first assembling and then annotating our 502 genomes using a pipeline of Prokka and Roary, we could identify those genes present in 99% of the genomes and define them as ‘core’, concatenating them into a core gene alignment. While recombination removal tools like Gubbins cannot be used alongside these programs, by only using the core genes, areas of high recombination and variation were not included in the alignment, effectively

'masking' them in the final phylogeny. We chose to use the total core genes, rather than any MLST-based methods, as the MLST approach loses much of the variation across the core genes by only focussing on the sequences of only a small number of highly conserved genes [136,137,340].

1,035 genes were designated as core by Roary; meaning they were present in 497 out of 502 of our genomes. 358 were designated as soft core (95 – 99%), 1,605 were designated as shell (15 – 95%), and 20,449 were designated as cloud (<15%). The total number of genes identified by Prokka annotation was 23,447. SNPs were called from this alignment using SNP-sites to create a phylogeny based on just the sites of variation within the core genes. This produced a final alignment 49,454 nucleotide bases long.

Taking this core gene SNP alignment, a phylogeny was created using IQ-TREE over 1000 pseudo-bootstraps [102]. In tandem with the phylogeny creation, the presence of antimicrobial genes within these genomes were assessed, and any candidate genes were identified using ARIBA [104]. The results of the AMR presence/absence testing, alongside temporal and spatial metadata including the country and decade of isolation, were plotted against the phylogeny in iTOL[112]. Figure 3.4 shows a graphical representation of this workflow.

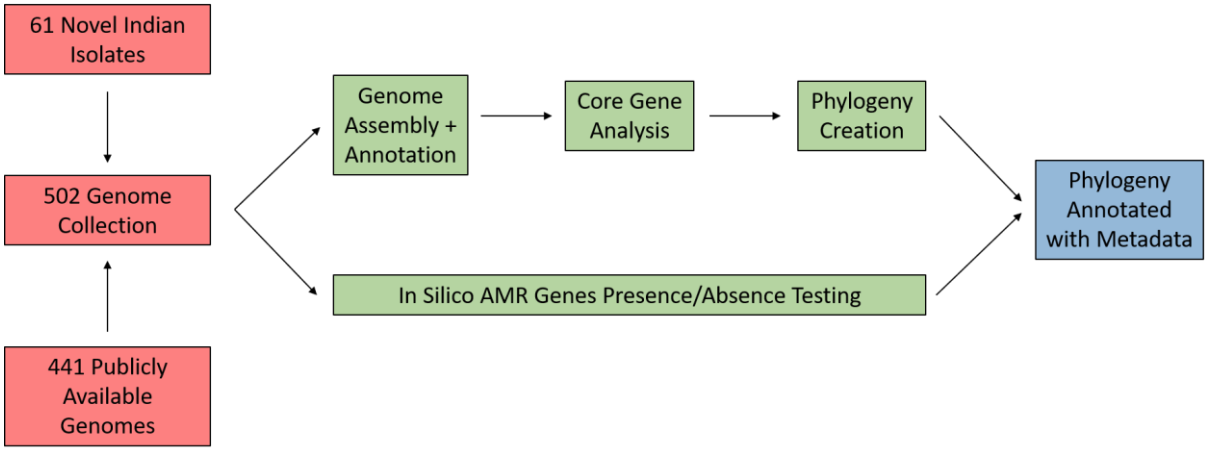

***Figure 3.24**: Flow chart diagram for generating the global phylogeny of C. diphtheriae and investigating the antimicrobial genes present in the genomes. Red boxes represent the data collection phase, green the data analysis phase, and blue the data interpretation phase.*

Figure 3.5 shows the core gene SNP phylogenetic tree of 502 *C. diphtheriae* genomes, annotated with the country and decade of isolation. Across the phylogeny multiple diverse clusters are present, contained within which are isolates from multiple countries and regions of isolation, as well as multiple decades of isolation. The divergence of these clusters indicates that *C. diphtheriae* has been established within the human population for at least a century and likely far longer.

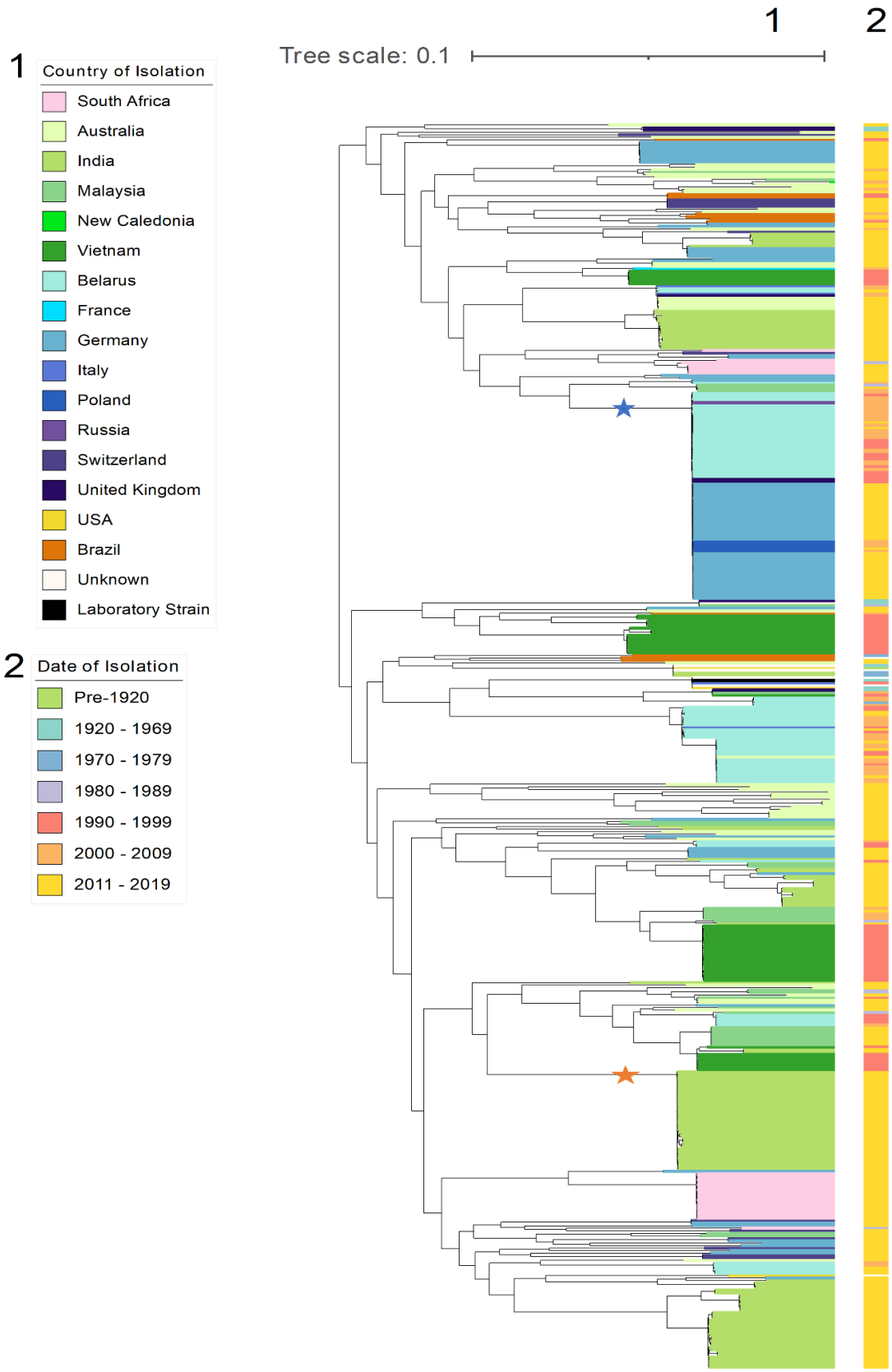

***Figure 3.25**: Maximum likelihood phylogenetic tree based on the extracted core gene SNPs from the 502 global C. diphtheriae genome collection. The country of isolation (1) and decade of isolation (2) are shown. Both the blue and orange*

stars highlight groups used for later BEAST analysis due to their close levels of inter-relatedness. The scale bar shows substitutions per site.

Within the major clusters, smaller monophyletic groups (closely related isolates with a recent shared ancestor) were clearly present. The vast majority of these monophyletic groups were isolated from a single country and within a relatively short time period, indicating a strong geographic and temporal association. An example of this type of cluster is marked by an orange star in Figure 3.5; a 40-isolate grouping with highly similar core gene SNPs, all isolated from India during 2011 – 2019 (in this case the four-year period 2015 – 2018). Within this group, 12 isolates were from Delhi, four were from Haryana, and 19 were from Uttar Pradesh in the North, while five were from Kerala in the South.

However, one monophyletic group notably did not follow this pattern. Marked by a blue star on the same Figure, the very large group dominated by German and Belarusian isolates showed incredibly conserved core genes, with very little SNP differences between them. Despite this, the genomes in this group were isolated across five countries (Germany, Belarus, Poland, UK, and Russia) over three decades, from as early as 1996 in Belarus, and persisting to at least 2017 in Germany.

Individual countries can be found represented across the phylogeny within multiple major clusters, and most commonly this includes countries in Europe and Asia. As an example, within India there are multiple monophyletic groups, such as the one marked with an orange star, isolated within the same four-year period (2015 – 2018) yet far apart on the core gene phylogeny. This shows that there are multiple distantly related and distinct clonal populations

circulating within this country at the same time causing disease. This pattern extends to other countries and regions, such as in Vietnam during the 1990s, and Belarus from the 1990s through to the 2010s, which both have multiple groups circulating concurrently and causing disease, despite the fact that they are only distantly related within the core gene phylogeny.

The MLST type was determined for each isolate, and each monophyletic group contained only one ST. Some isolates within these groups did however register as 'novel ST' using MLSTcheck. 100 MLST types were represented across the collection, with 73 isolates classified as 'novel ST'.

3.4 Antimicrobial resistance in *C. diphtheriae*

Figure 3.6 shows the global phylogeny of *C. diphtheriae* with the presence and absence of AMR genes identified using ARIBA plotted alongside the country and decade of isolation. The number of AMR genes present within the 502 genomes was generally higher in those isolated more recently, especially among those isolates from the 2010s, and the same can be said for the variety of resistance type classes that these genes represent. This is especially true in India, where many of the most recent isolates are from. Additionally, similar profiles of AMR genes to those categorised from India are present in a small number of recently isolated genomes from Germany and Switzerland. While most countries and regions showed highly variable AMR gene profiles across their genomes, Indian isolates showed a much more conserved profile. This is despite the large distances these monophyletic groups are apart within the phylogeny.

Sulphonamide resistance conferred by the resistance gene *sulI* was the most common AMR determinant across the 502 genome collection, consistently reported in the Indian isolates as well as sporadically found among other Asian and European genomes, most notable in Vietnam, Germany, and Belarus. Genes conferring resistance to Aminoglycosides (*aadA4, aph3_Ia, strA* and *strB*), chloramphenicol (c*mr*) and trimethoprim (*dfrA1*) were present to differing degrees in almost all India genomes, as well as in isolates from Germany, Switzerland, and Vietnam. Only one isolate across our whole collection (Isolated in India in 2015) was found to contain a macrolide resistance-encoding gene (*msrA*) while no genes conferring resistance to β-lactamases were found. This is despite the fact that erythromycin and penicillin are the traditionally recommended antibiotics for treating confirmed cases of early-stage diphtheria. Phenotypic AMR testing was carried out for the 61 novel Indian isolates and these results align well with the genes identified *in silico*.

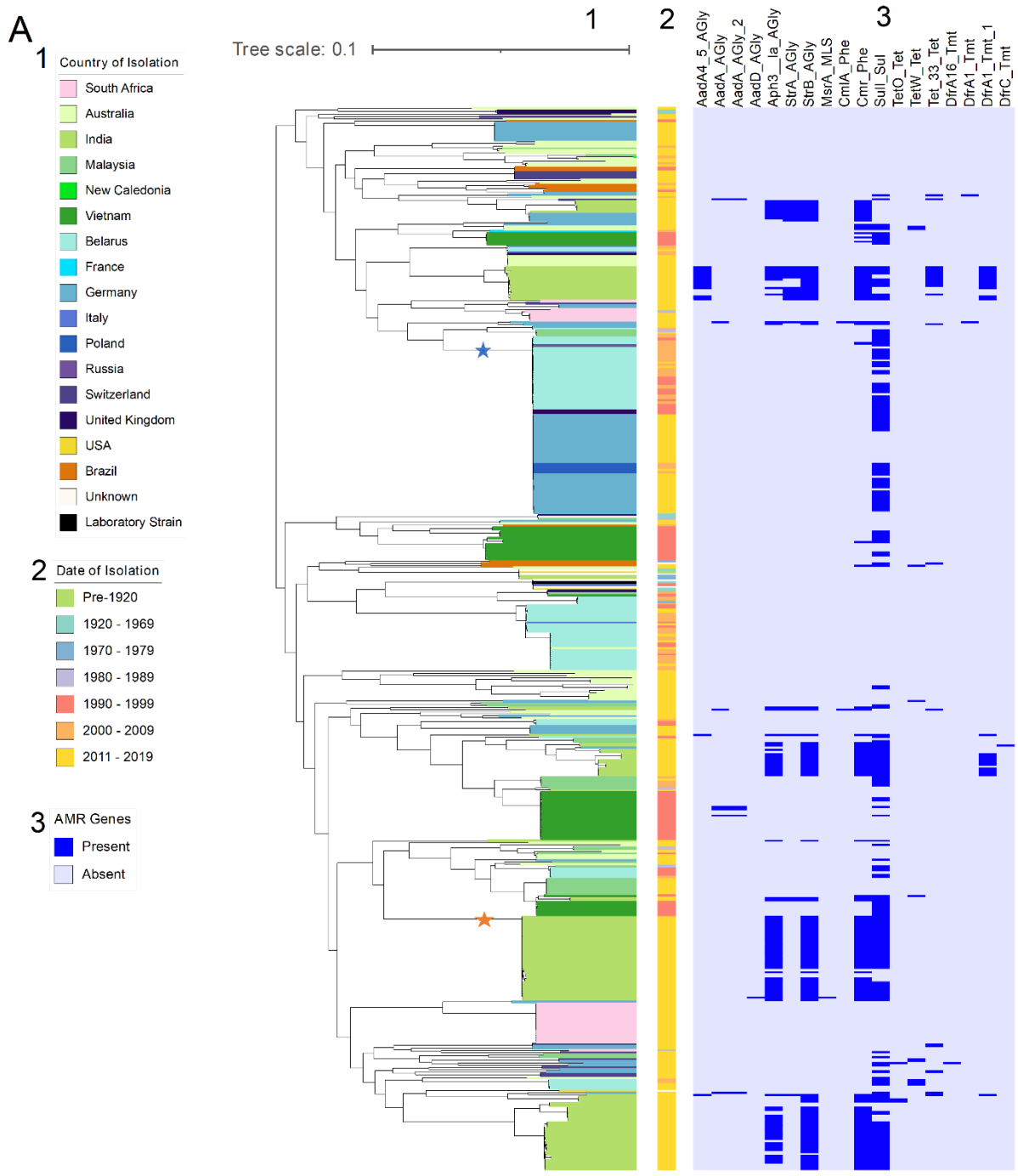

Figure 3.26: *Maximum likelihood phylogenetic tree based on the core gene SNPs from the 502 global C. diphtheriae genome collection. Both the blue and orange stars highlight groups used for later BEAST analysis due to their close levels of inter-relatedness. (3) shows the presence (dark blue) and absence (light blue) of AMR genes as a heatmap, identified using ARIBA. The scale bar shows substitutions per site.*

Due to the heavy temporal bias towards recent decades, especially the 2010s, we averaged the number of AMR genes present within genomes isolated in the same decade. By doing this, we aimed to create a more balanced view of AMR genes across the decades of isolation. Even after doing so, the 2010s continued to show the highest number of AMR genes per genome, almost four times higher than the next highest decade; the 1990s. Figure 3.7 shows the average number of AMR genes per decade, coloured by the class of antibiotic the genes confer resistance to. There was a significant positive correlation between the average number of AMR genes per genome, and the decade in which those genomes where isolated using a Pearson's product-moment correlation test ($r\ (9) = 0.68$, $p = 0.02$).

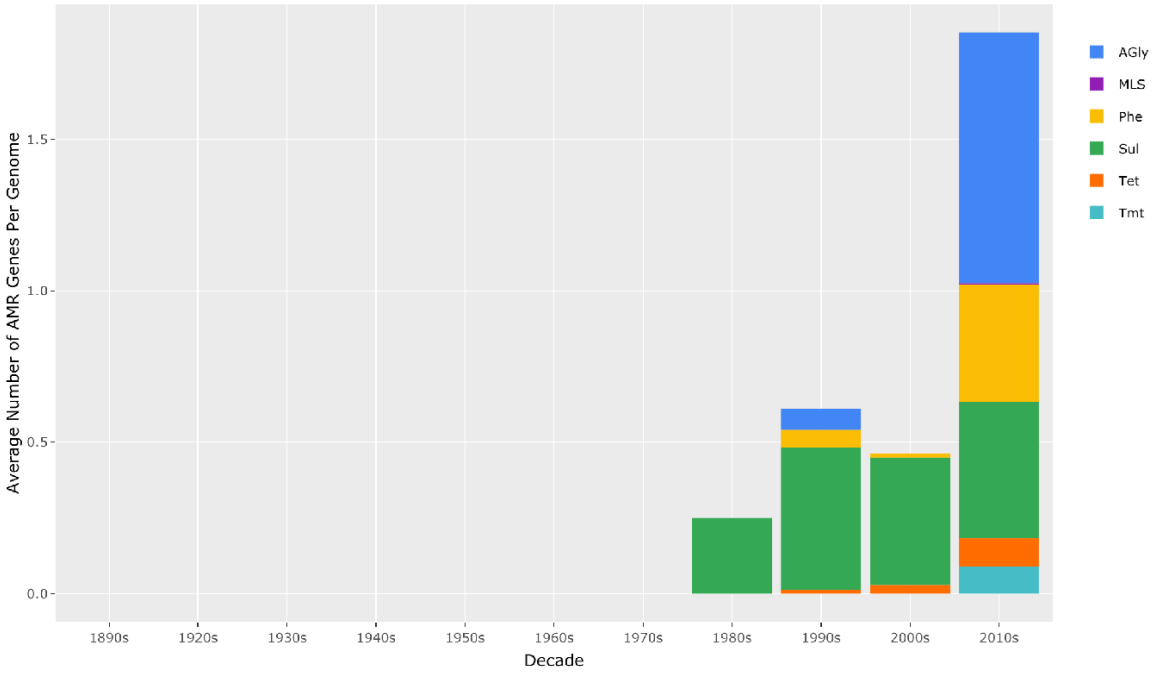

Figure 3.27: *Antimicrobial resistance by decade. The coloured bars represent the average number of genes per genome found that represent each class of antibiotic per decade. AGly (blue) = aminoglycosides, MLS (purple) = macrolide-lincosamide-streptogramin, Phe (yellow) = phenicols, Sul (green) = sulfonamides, Tet (orange) = tetracyclines, Tmt (light blue) = trimethoprim.*

The 2010s also showed the greatest variation in the number of classes of antibiotic represented by resistance determinants, resulting in six classes. By comparison, the 2000s presented genes that confer resistance to three classes and the 1990s showed resistance to four. The 1980s meanwhile presented genes that confer resistance to only one class (sulphonamides), and no AMR genes were detected by ARIBA in genomes from the prior decades. Indeed, no AMR genes were detected in genomes isolated in the decades prior to the 1980s.

3.5 Time scaled phylogenetic analysis

Due to the challenges faced with mapping the highly variable collection of *C. diphtheriae*, an accurate time scaled phylogeny using BEAST was not feasible. BEAST analysis requires mapped alignments that have had recombination signatures removed, and as mentioned in 3.2, neither mapping the whole collection to a single reference nor recombination removal tools could accurately work on such a diverse and recombination-heavy genome collection. A time scaled phylogeny is however an important tool in understanding the evolution of a population of a pathogen, adding crucial estimates of dates for branching lineages, introduction events and transmission dynamics.

In an attempt to address this problem, we chose two large monophyletic groups containing highly similar isolates based on their core genes and mapped them to the closest completed genome available. The first was the large Belarus- and Germany-dominated European group marked with a blue star in Figure 3.5, which was mapped to the complete genome NCTC 13129. The second was the Indian group marked by an orange star in Figure 3.5, which was

mapped to NCTC 11397. Gubbins was then used to remove recombination signatures from these mapped alignments, before the now-recombination free alignments were used as the basis for BEAST time scaled phylogenetic analysis. In the European group, the ratio of pre- and post-recombination removed SNPs included in the alignment was 1.1 (1944:1783), while in the India group the ratio of pre- and post-recombination removed SNPs included in the alignment was 1.2 (26879:22717).

Figure 3.8 shows the time scale phylogeny for the European group, and Figure 3.9 shows the Indian group time scale phylogeny. Within the European group, three clades are present, the largest of which spans 1987 – 2008. The second and third clades are estimated to have diverged in July 1985 with the former present from September 1988 to 2010 and the latter estimated to have been in circulation between April 1990 and 2017. The most recent shared ancestor of all the isolates within the group was estimated to have been in September 1983.

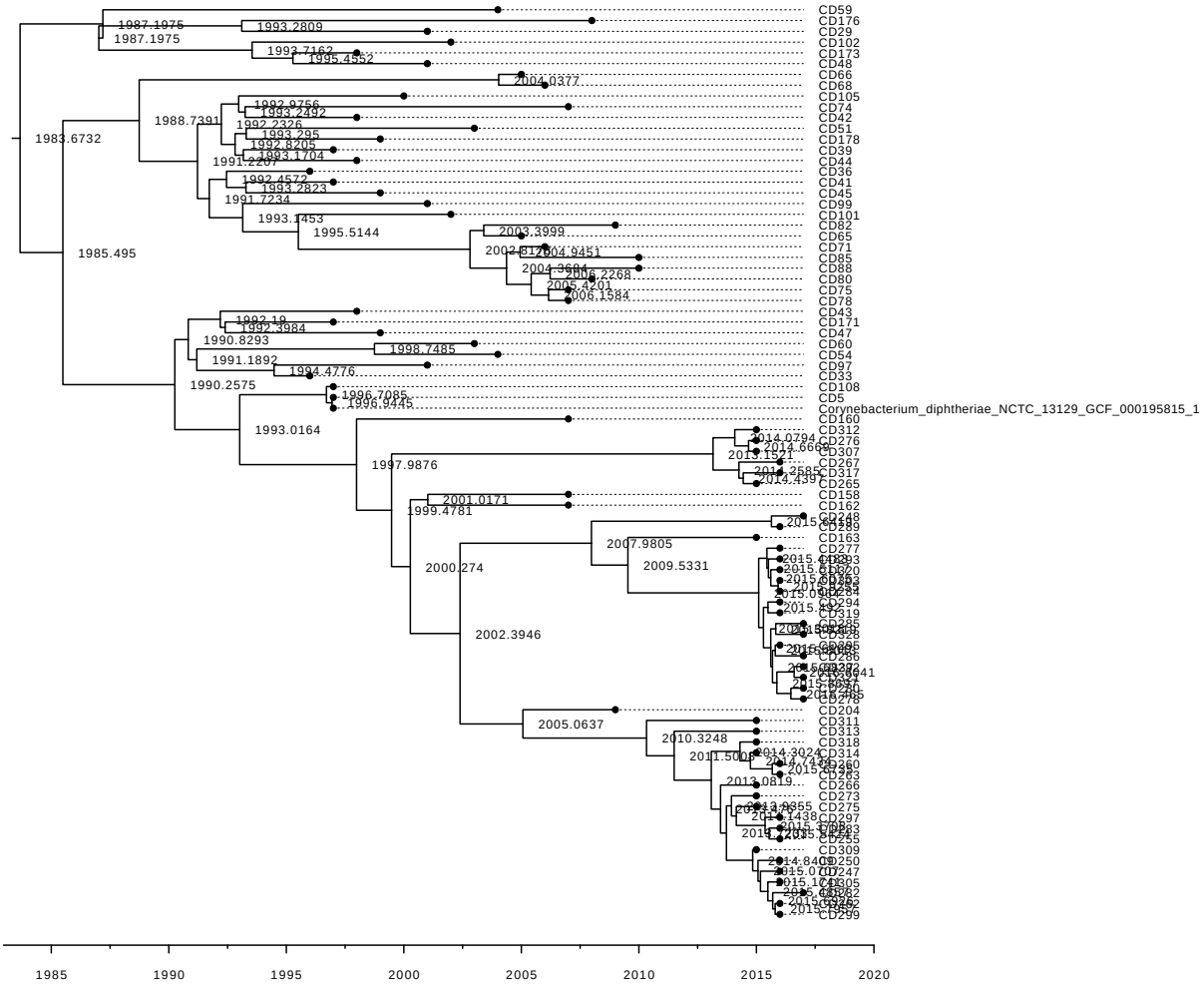

Figure 3.28: *BEAST phylogeny of the large monophyletic Belarus- and Germany-dominated European group, marked by a blue star in Figures 3.25 and 3.26.*

Analysis of the Indian group shown in Figure 3.9 revealed two clades, which were estimated to have diverged in February 2009. The first clade was estimated to have been present from April 2012 to 2017, while the second clade was estimated to have been present from June 2013 to 2018. The distance between NCTC 11397 and the rest of the isolates was much larger than between NCTC 13129 and the European group genomes. The shared ancestor between the Indian isolates and NCTC 11397 was estimated to have been in October 1955.

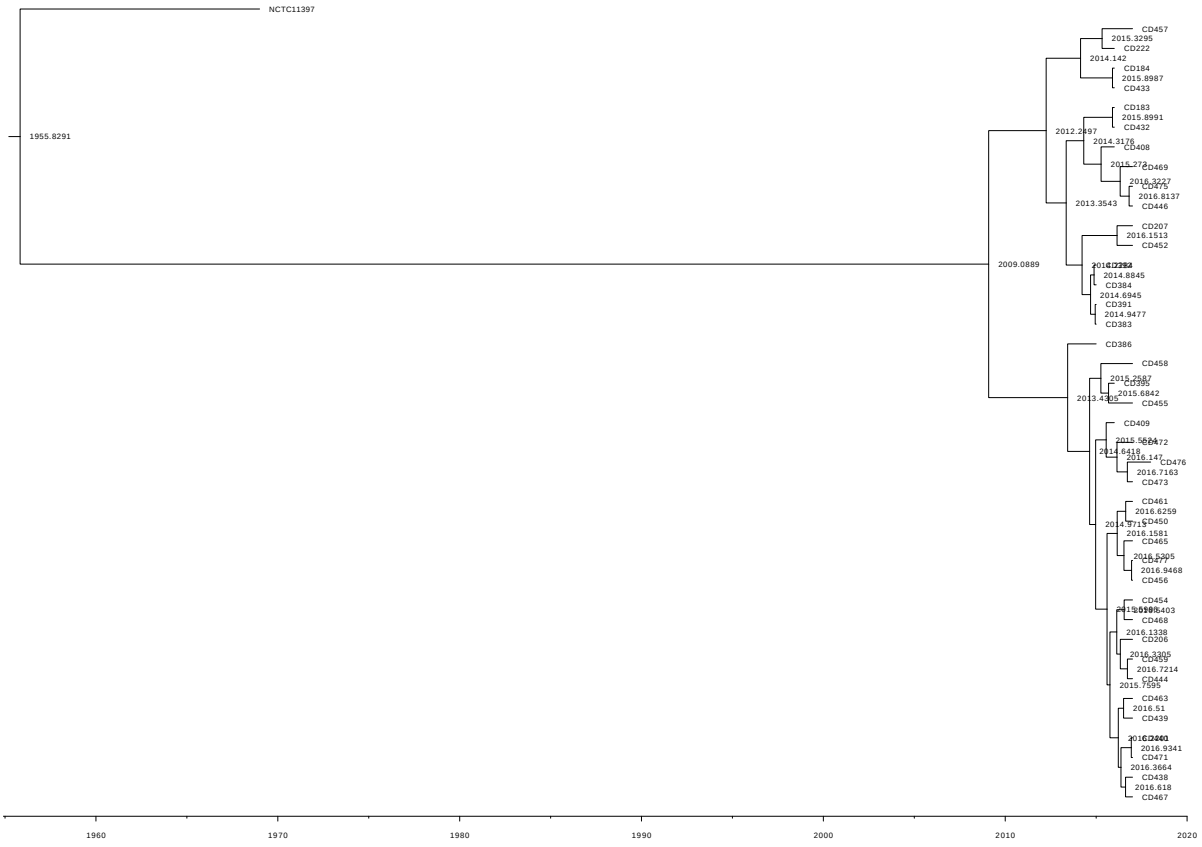

Figure 3.29: *BEAST phylogeny of the monophyletic Indian group marked by an orange star in Figures 3.25 and 3.26.*

3.6 Creating a national collection of *C. diphtheriae*

By taking our 61 novel genomes from India and combining them with 61 publicly available genomes, a 'national' collection was created. The 122 genomes covered 6 states with 22 from Delhi, seven from Haryana, two from Himachal Pradesh, and 35 from Uttar Pradesh in the North of the country. Additionally, 31 were from Kerala and 25 from Tamil Nadu in the South. The two isolates from Himachal Pradesh were isolated in 1973, while all other isolates were isolated from 2015 – 2018: nine from 2015, 72 from 2016, 37 from 2017 and two from 2018. Figure 3.10 and Figure 3.11 show the states and years of isolations.

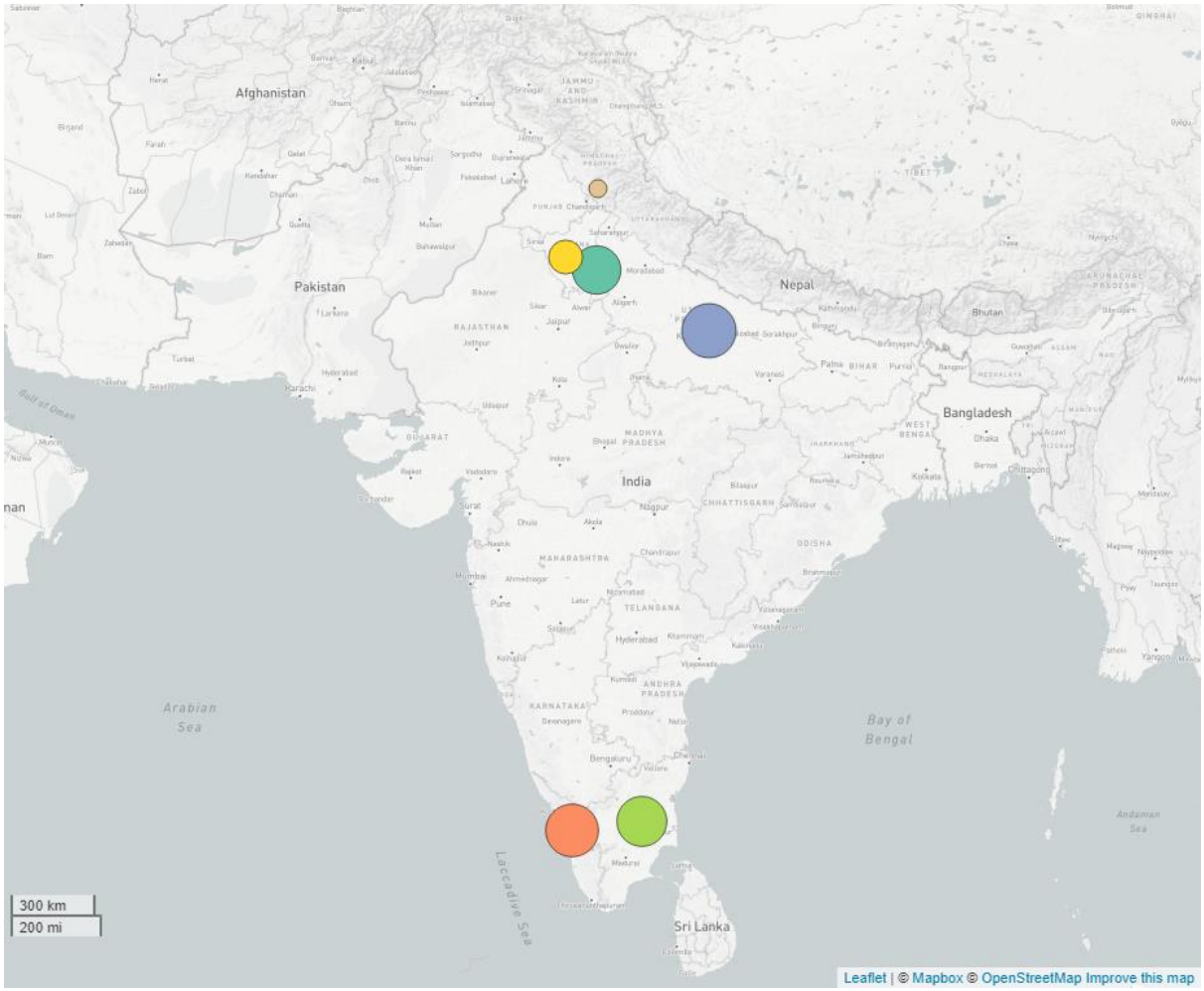

***Figure 3.30**: Map of 122 C. diphtheriae genomes, coloured by state of origin and scaled by number of isolates. Peach = Himachal Pradesh, Yellow = Haryana, Turquoise = Delhi, Blue = Uttar Pradesh, Orange = Kerala, Light Green = Tamil Nadu.*

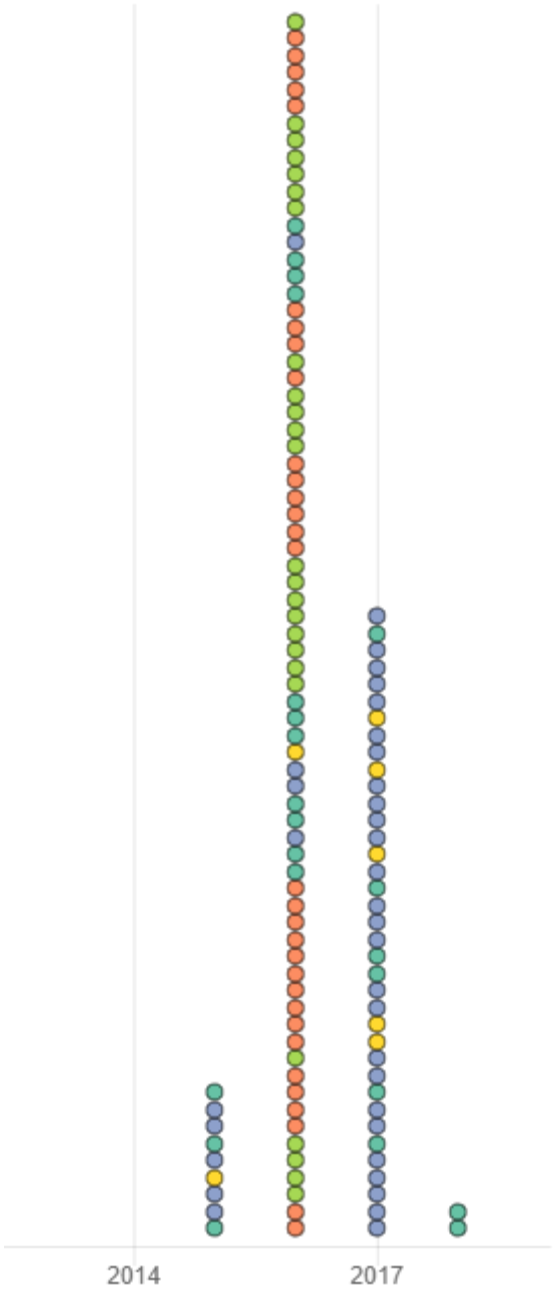

Figure 3.31*: Timeline of 120 C. diphtheriae genomes, coloured by State of origin. Two isolates from 1973 isolated in Himachal Pradesh were also included in the collection.*

Of the novel 61 genomes we sequenced, all were isolated in Northern India including all the 22 isolates from Delhi, all seven from Haryana, and all 32 from Uttar Pradesh. These novel genomes represented isolates covering the four-year period 2015 – 2018 with the exception

of the isolate from Himachal Pradesh in 1973. All nine of the 122 isolated in 2015, 16 of the 72 isolated in 2016, 34 of the 37 isolates in 2017, and both 2018 isolates belonged to the 61 novel genome collection.

After these genomes had been collected, we applied the same methodology used for the global collection. By using the exact same methodology, in addition to ascertaining the population structure of *C. diphtheriae* within India, this allowed direct comparison of the results between this national and the international collection.

3.7 An Indian phylogeny for *C. diphtheriae*

Following the same methodology as used for the global collection, the 122 Indian genomes were collated by combining 61 novel genomes and 61 genomes from publicly available sequences. The assembled genomes were annotated using Prokka, before core gene analysis using Roary, with the same cut-offs. A gene must be present in 99% of genomes (121 out of 122) to be considered core. These core genes were concatenated together for each genome and aligned before SNPs were called. This SNP alignment was then used to create the phylogeny using IQ-TREE over 1000 pseudo-bootstraps. Simultaneously, the presence of antimicrobial genes within these genomes was once again assessed, and any candidate genes were identified using ARIBA. The results of the AMR presence/absence testing, alongside temporal and spatial metadata including the state and decade of isolation, were plotted against the phylogeny in iTOL. Figure 3.12 shows the graphical representation of this workflow adapted for the Indian collection.

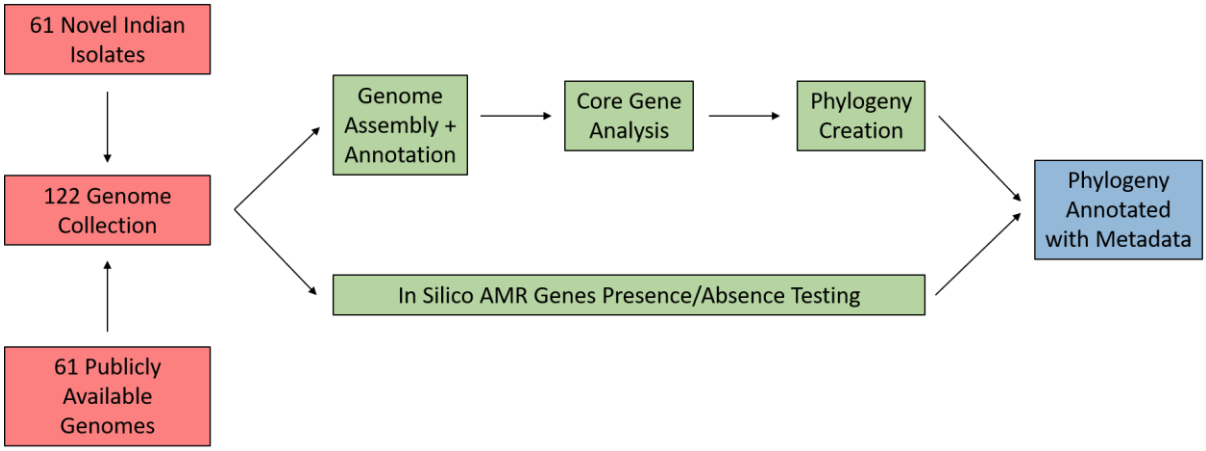

Figure 3.32: *Flow chart diagram for generating the Indian national phylogeny of C. diphtheriae and investigating the antimicrobial genes present in the genomes. Red boxes represent the data collection phase, green the data analysis phase, and blue the data interpretation phase.*

Within the Indian collection, 1,367 genes were designated as core, higher than the 1,035 identified in the global collection. 196 were designated as soft core, 1,465 were designated as shell, and 4,408 were designated as cloud genes. This compares to the 358, 1,605, and 20,449 identified respectively in the global collection. The total number of genes in the pan gene list across all 122 genomes was 7,436, much lower than the 23,447 identified in the global collection. Figure 3.13 shows the gene number breakdowns for both the global and Indian collections.

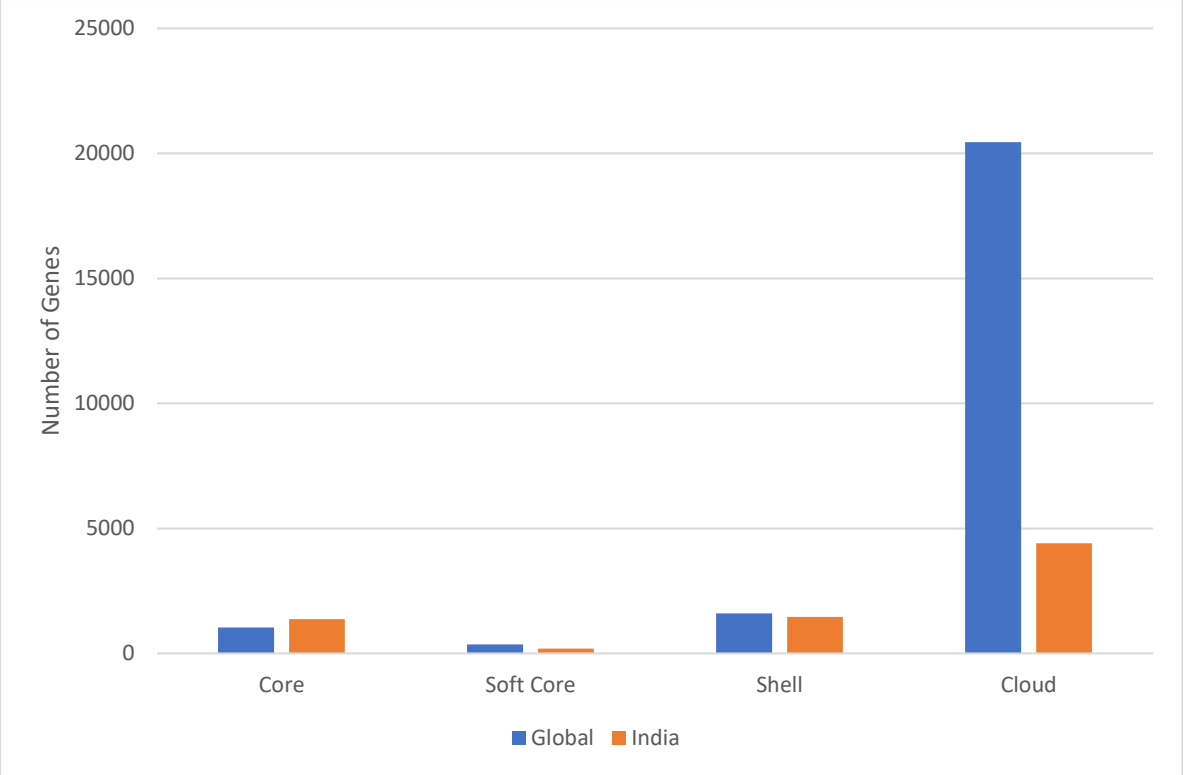

Figure 3.33: *Genes identified by Roary and Roary Plots, with the global collection in blue and the Indian subset in orange.*

Figure 3.14 shows the core gene phylogeny of the 122 Indian *C. diphtheriae* isolates. Smaller monophyletic groups harboured isolates from individual states and common years of isolation, mimicking the pattern observed across our global phylogeny. Within those monophyletic groups isolated across multiple states, the majority were distinct to either those states from Northern India – Delhi, Haryana, and Uttar Pradesh – or those in the South of the county – Kerala and Tamil Nadu. The single shared year of isolation was also a feature of these grouping. Within the larger monophyletic groups however, such as the one marked by a purple star, genomes represented isolates from across the majority of states, including isolates from both Northern and Southern states, as well as across all four of the years of isolation spanning 2015 – 2018.

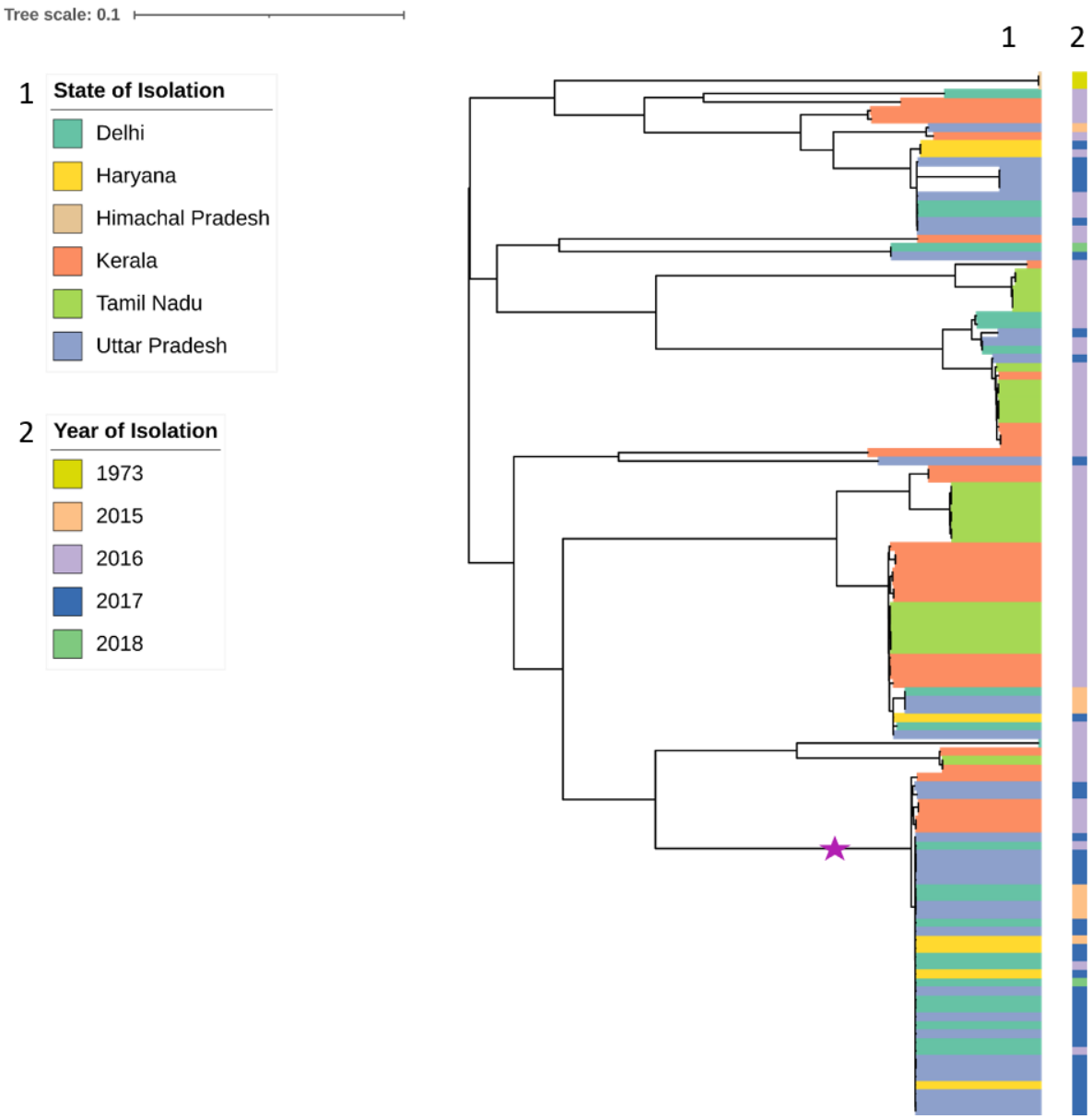

***Figure 3.34**: Maximum likelihood phylogenetic tree based on the extracted core gene SNPs from the 122 Indian C. diphtheriae genome collection. The state of isolation (1) and year of isolation (2) are shown. The scale bar shows substitutions per site.*

This pattern largely mimics the global population of *C. diphtheriae*. Individual monophyletic groups share a close geographical origin – in this case single or neighbouring state level rather than country level. In some cases, though, closely related isolates have apparently spread more successfully to cover a much larger geographical area. The group marked by a

purple star in Figure 3.14 shows similarity to the large European clade marked by a blue star in Figure 3.5, as isolates successfully spread to cover a large geographical area. Although the four years represented here is a much smaller time period than the three decades of the blue star group in Figure 3.5, it is still a distinctly more complex feature within this state level phylogeny, as all other monophyletic groups were isolated from a smaller time frame. This reinforces the concept that highly similar clonal *C. diphtheriae* strains can remain conserved within distant populations for extended periods of time.

Figure 3.5 also showed than within countries, including India, multiple distantly related monophyletic groups of *C. diphtheriae* circulate at the same time, causing disease concurrently. This pattern continues to the state level, where monophyletic groups isolated from the same individual states were found multiple times throughout the phylogeny. This was true among both our novel genomes from the North and public genomes from the South, as groups of *C. diphtheriae* isolated within the same state and the same year were often only distantly related within our phylogeny.

3.8 Antimicrobial resistance of *C. diphtheriae* within India

As mentioned in section 3.3, the number of AMR genes present across the global 502 genome collection increased in recent decades, and as India makes up almost 40% of the 2010 isolates, the AMR determinants present within our 122 genome Indian collection is relatively high. Figure 3.15 shows the Indian phylogeny of *C. diphtheriae* with the presence and absence of AMR genes identified using ARIBA plotted alongside the country and decade of isolation. As with the global phylogeny in Figure 3.6, isolates from the same individual

monophyletic groups mostly share the same pattern of AMR genes present or absent within their genomes.

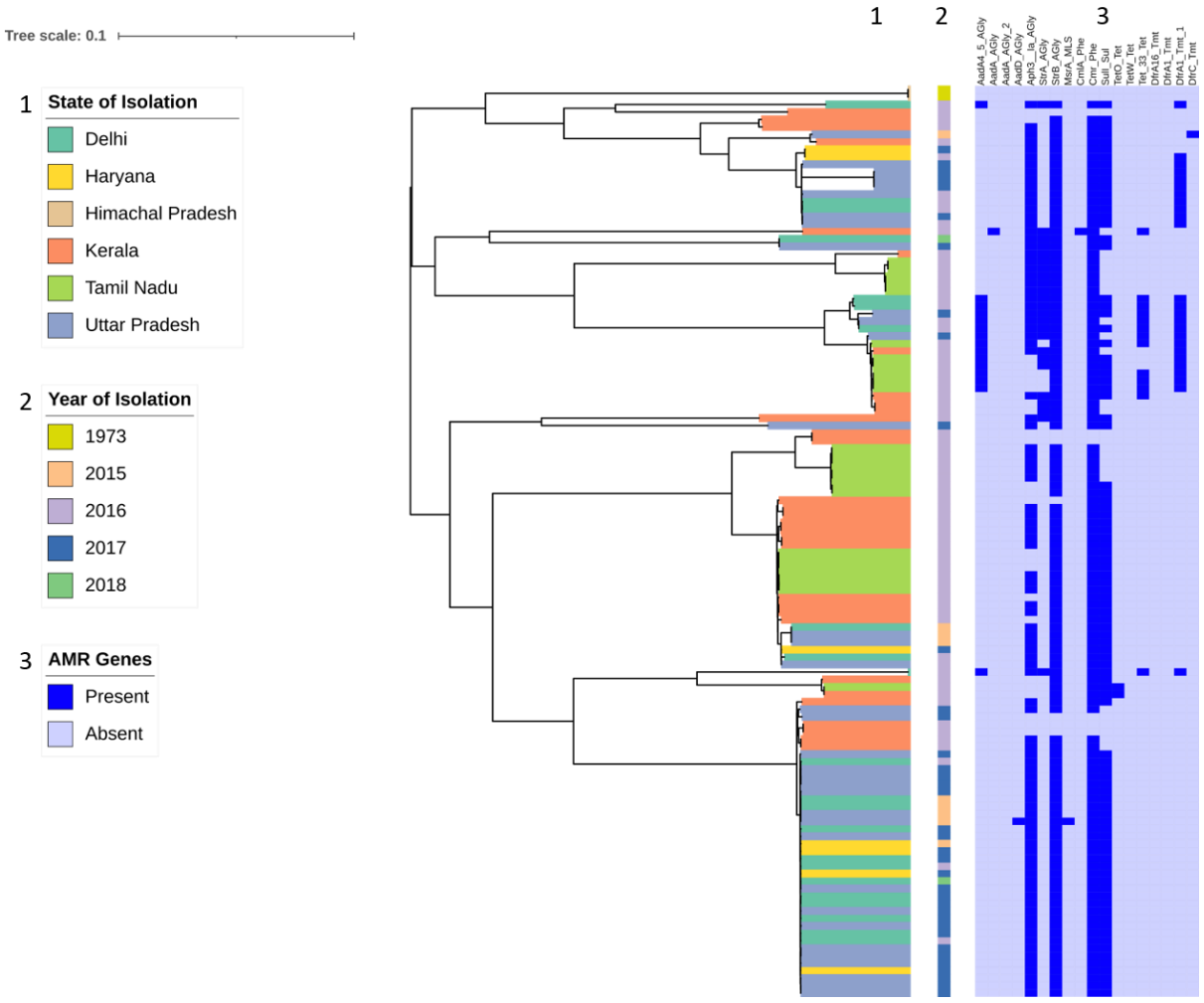

***Figure 3.35**: Maximum likelihood phylogenetic tree based on the extracted core gene SNPs from the 122 Indian C. diphtheriae genome collection. The state of isolation (1) and year of isolation (2) are shown. (3) shows the presence (dark blue) and absence (light blue) of AMR genes as a heatmap, identified using ARIBA. The scale bar shows substitutions per site.*

Genes conferring resistance to chloramphenicol, aminoglycoside, and sulphonamide classes were most common across the 122-genome collection, with Northern states harbouring higher

levels of AMR genes within their genomes than those from Southern states. This included the only isolate to have a macrolide resistance determinant identified by ARIBA, in Uttar Pradesh. Figure 3.16 shows a state-level breakdown of the classes of antibiotic AMR genes found within genomes isolated from each state.

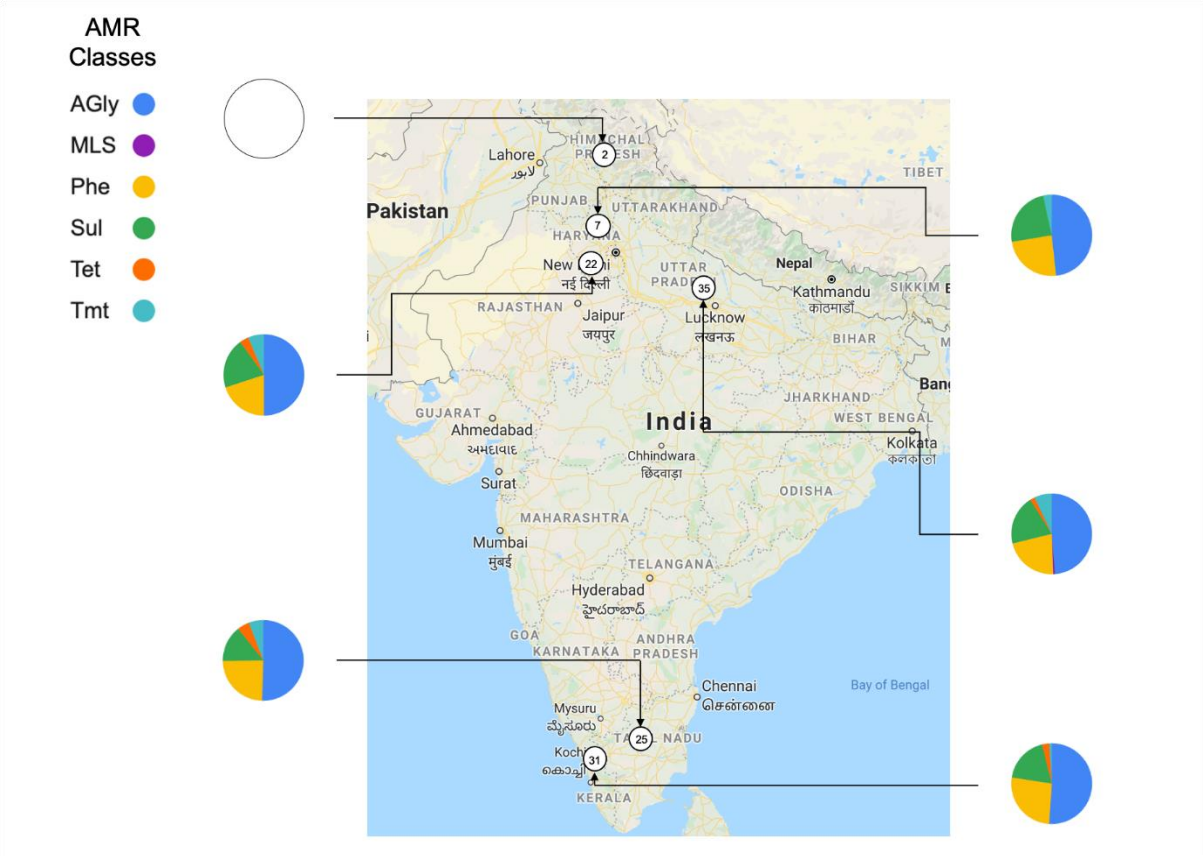

***Figure 3.36**: The AMR gene proportions represented in the 122 C. diphtheriae genomes from India at state level. The number of isolates from the six states sampled are shown in circles on the map. The AMR genes are coloured by the classes of antibiotic the genes offer resistance to. AGly (blue) aminoglycosides, MLS (purple) macrolide–lincosamide–streptogramin, Phe (yellow) phenicols, Sul (green) sulfonamides, Tet (orange) tetracyclines, Tmt (light blue) trimethoprim. The map was taken from Google Maps* [341].

The two genomes for isolates from Himachal Pradesh in 1973 carried no detectable AMR genes. Among the other 120 Indian genomes isolated from 2015 – 2018, the 7 isolates from

Haryana were found to harbour genes conferring resistance to 4 classes of antibiotic (aminoglycosides, phenicols, sulphonamides, and trimethoprims). The 22 isolates from Delhi, 25 from Tamil Nadu and 31 from Kerala were found to harbour 5 (aminoglycosides, chloramphenicols, sulphonamides, trimethoprims and tetracyclines) and Uttar Pradesh's 35 isolates harboured 6; the same 5 of aminoglycosides, chloramphenicols, sulphonamides, trimethoprims and tetracyclines, in addition to the one macrolide resistance gene-positive isolate. As stated previously, no β-lactamase resistance genes were identified using ARIBA.

3.9 Discussion

By sequencing 61 novel genomes and combining them with 441 publicly available genome sequences, we were able to gain a more comprehensive appreciation of the population structure of *C. diphtheriae* around the world. We identified several large phylogenetic clusters, each containing numerous smaller monophyletic groups from multiple countries covering decades of isolation. Isolates from individual counties were frequently present throughout the phylogeny and across multiple clusters, suggesting that numerous diverse sublineages of *C. diphtheriae* are successfully coexisting, all circulating within the geographical population. This situation is somewhat distinct to that found in some other bacterial pathogens, where a small number of highly clonal lineages dominate after having spread across the globe. Such an example is *Mycobacterium abscessus*, an emerging Gram-positive bacterium that causes nosocomial respiratory infections [342].

In a national context, a very similar picture is present. Across India, individual monophyletic groups are usually state or neighbouring state-specific and from a single year of isolation.

However, the large clade marked by a purple star in Figure 3.14 shows that some clones of *C. diphtheriae* persist across both Northern and Southern states, and across multiple years of isolation (in this case four years from 2015 – 2018). More regular isolations and sequencing of *C. diphtheriae*, both historically, and moving into the future, will help to present a much more defined picture of diphtheria in countries like India, where the prevalence and risk of diphtheria infection and transmission are much higher.

The large European group marked by a blue star in Figure 3.5 (and to a lesser extent the purple star denoted group in Figure 3.14) raises an important question around re-emergence. Across the group there is only minimal difference in the core genes, yet the genomes were generated from isolates collected over three decades in multiple different countries. Due to the lack of evidence of *C. diphtheriae* reservoirs existing within animal populations, or in the environment, asymptomatic human carriage could be one potential answer for the continued re-emergence of highly similar isolates across time. The large genetic distances between isolates also add more evidence to this theory, as reported clinical diphtheria case numbers do not link to sufficiently regular outbreaks to sustain the clonal bacterial population over such long periods. Asymptomatic carriage is thus a potential explanation for the persistence of lineages.

In areas of high vaccine coverage, as is the case in much of Europe, the potential for incompletely protected individuals to act as a reservoir for *C. diphtheriae*, allowing the disease to quickly re-emerge under favourable circumstance, clearly requires further investigation. An important note to add though is the relative scale of countries in Europe compared to India, which is almost as big as much of Europe itself. Interestingly, citizens of

European countries can move relatively freely across much of the continent for both business and leisure and this might facilitate spread. Geographical borders do not act as much of a deterrent to disease spread in the modern world.

The increase in the frequency of carriage of resistance determinants found in recent decades goes hand-in-hand with trends observed in many other pathogens ([252]). Studies investigating a broad spectrum of microorganisms have reported that AMR has been increasing, most notable in the last few decades. Additionally, the AMR data for *C. diphtheriae* could be an indication of more evidence for human carriage, where indirect exposure to antibiotics could play a role. The lack of resistance determinants for macrolide resistance and β-lactamases production was perhaps surprising, given the position of erythromycin and penicillin as the first line recommended antibiotic in the treatment of diphtheria cases (alongside antitoxin). In contrast, the prevalence of genes conferring resistance to classes not recommended for the treatment of diphtheria cases was intriguing.

The similar AMR profiles exhibited in recent Asian isolates, as well as sporadically in recent isolates from Europe, could suggest that this development of resistance has been driven by the indirect exposure to antibiotics, which in turn has driven the uptake of mobile elements harbouring resistance elements to multiple different classes of antimicrobials. This could also be an example of 'collateral selection' (also termed 'collateral resistance'), where pressure to acquire resistance against any one antimicrobial agent drives the advancement of resistance to other agents [343]. The ability of *C. diphtheriae* to recombine DNA into its genome, alongside co-habitation of the lower human throat with other species in the microbiome provide a niche for *C. diphtheriae* to gain access to novel genes, including those that confer AMR. Potential

exposure to a wide variety of antibiotics not traditionally associated with the treatment of diphtheria could add more evidence to the theory of asymptomatic human carriage. By colonising the lower throat of a human, *C. diphtheriae* could be exposed to antibiotic treatments targeting other diseases. This is especially plausible in areas where antibiotics can be obtained 'over the counter'.

The high levels of diversity and recombination in *C. diphtheriae* highlights the difficulty in using traditional mapping-based methods, as these factors heavily impacts the robustness and confidence in any results produced. The core gene methodology still remains the most effective way to create accurate and robust phylogenies of *C. diphtheriae*, as this provides a way to counter the high diversity and levels of recombination across the species. By only building trees based on those genes present in large percentage of genomes, this allows areas of high recombination or diversity to be masked, as they would otherwise negatively impact the phylogeny creation. Despite all this, we have shown that mapping-based approaches can be effective when used on smaller, more closely related isolates from within an individual monophyletic group. By choosing the nearest complete genome, rather than a distant reference, fastq read files can be successfully mapped. Recombination removal tools such as Gubbins can also be effective when used under these circumstances, allowing the use of BEAST to create time scaled phylogenies and make estimations for key dates. These offer new ways of approaching individual outbreaks of diphtheria and opens up the possibility of using time scaled phylogenies to better understand individual clonal outbreaks. These tools can also provide strong support for public health measures, such as an estimation of the date of introductions, as well as dating common ancestors and clade branching events.

There are caveats to this method, however. Any outbreak would have to be caused by only a single closely related group of *C. diphtheriae* bacteria, rather than the numerous distantly related clusters causing infection concurrently as is the case in both India and elsewhere. Another caveat is that there are not, as yet, enough high-quality finished genomes available to accurately act as a reference for all parts of the global population, and this reaffirms the importance of continuing to create them. As sequencing technologies improve and become more accessible, long read data and hybrid assemblies will allow many more opportunities for these kinds of analyses on diphtheria outbreaks caused by *C. diphtheriae*.

While we have presented theoretical evidence on the potential for asymptomatic *C. diphtheriae* carriage, our study only includes genomes from published clinical cases. As *C. diphtheriae* is not routinely screened for in otherwise healthy individuals, and the lower throat where the bacterium resides is not swabbed routinely during check-ups due to the invasiveness of the procedure, larger collections of carriage isolates do not exist at this time. However, the spatio-temporal structure of the phylogeny suggests that asymptomatic carriage could be playing a very important role in the life cycle of *C. diphtheriae*, as highly similar bacteria continue to re-emerge repeatedly over decades, with seemingly large time spans in between outbreaks. This begs the question, 'where are they hiding'? Carriage would also present an opportunity for the uptake of AMR genes and evolution driven by recombination that has been shown to be so typical of the species. Future studies based around monitoring human populations, especially in at-risk communities, in both vaccinated and non-vaccinated individuals is key to further investigating and understanding the potential of carriage and could provide an answer to the increase in AMR, especially over the past decade.

The rising number of diphtheria cases in recent years show the potential for widespread re-emergence of the disease under the right circumstance. Both in Yemen and among the Rohingya refugee camps, a large amount of people confined together with a lack of access to healthcare, sanitation, and low vaccine coverage created an opportunity for *C. diphtheriae* to spread and infect thousands [222] . With 2019 the highest number of reported cases to the WHO in 23 years, these optimal opportunities for *C. diphtheriae* or 'perfect storms' are becoming more common. While treatment of *C. diphtheriae* infection seemingly remains unchanged, COVID-19 has negatively impacted the vaccination schedules of many children across the world as families move away from cities in an attempt to avoid the virus. This is especially true in areas where the risk of catching diphtheria is higher. This could have a knock-on effect for years to come, as pockets of children do not acquire protective immunity to these preventable diseases. it is arguably more important than ever to understand this historically disease, and to prevent it from becoming a major global threat ever again in its original, or a modified, better adapted, form.

4.0 National and global diversity of the *C. diphtheriae tox* gene and the diphtheria toxin

4.1 Introduction

The diphtheria toxin (DT) was first reported to be the active disease component in cases of diphtheria over a century ago [128,196]. Only five years after Edwin Klebs proved that *C. diphtheriae* was the causative agent of the disease, and four after Friedrich Löffler first cultured the bacterium, Émile Roux and Alexandre Yersin showed that it was products produced by *C. diphtheriae*, rather than the bacterium itself, that caused symptoms [128,194–196]. The elucidation of DT as the product in question lead to the development of the prototype diphtheria vaccine, a toxoid formulation designed to provide an inactivated DT for the body to produce antibodies against. Despite the vaccine being a toxoid formulation, some *C. diphtheriae* cellular material is carried over as a part of the production process, potentially offering some level of protection against the bacterium itself, as well as the secreted toxin [220]. The vaccine is part of most national immunisation programs, with doses given in the early weeks post birth. This can be followed by boosters in later years for those at risk.

The *tox* gene itself is carried on a corynephage, and post-infection the phage and its gene content can become incorporated into the *C. diphtheriae* genome [204]. Non-toxigenic *C. diphtheriae* causing disease have been reported in increasing numbers over recent years, and while the exact pathogenesis of how these isolates cause symptoms is not yet fully understood, they displayed the capacity to cause invasive or systemic infections [135,205–207].

Non-toxigenic isolates are primarily those that lack the *tox* gene entirely, although non-toxigenic toxin gene bearing (NTTB) isolates, where the *tox* gene is present but has become defunct, have also been reported [135,205–207]. The corynephage is lytic in some species closely related to *C. diphtheriae*, including *C. ulcerans* and *C. pseudotuberculosis*. These species have been associated with the ability to cause diphtheria-like infections, usually through the expression of the diphtheria toxin protein. Both *C. ulcerans* and *C. pseudotuberculosis* are zoonotic diseases, and there has been an increase in reported case numbers in recent years [193,202,344].

DT is part of the AB toxin family, a group that also includes tetanus toxin, anthrax toxin, and cholera toxin, although cholera toxin belongs to a specific subclass of the family known as AB_5. The term 'AB', as discussed in section 1.5 of the Introduction and shown in Figure 1.16, is due to the dual-subunit structure of the toxin released by the cell after synthesis. DT was the first toxin of this group classified. The B-subunit binds with target receptors on a host cell and facilitates transference across the lipid bilayer. Once inside a target cell, the bonds keeping the A and B subunits together are broken, activating the toxic activity of the A subunit, a process which can lead to cell death.

The crystal structure of DT, a Y-shaped molecule of three domains, was first published to 2.5 Å resolution by Choe *et al* in 1992 [345]. The toxin is composed of a catalytic domain (C) (the constituent part of subunit A), a transmembrane (T) domain, and the receptor-binding (R) domain, with T and R making up the B subunit. The receptor that the B subunit binds to on human cells is the heparin-binding epidermal growth factor (HB-EGF). Once inside a cell, the disulphide bridge and peptide bond holding the A and B subunits together breaks. This

dissociation activates subunit A, which targets and inactivates eukaryotic elongation factor 2 (EEF2), preventing protein synthesis within the cell [238,346,347]. By inhibiting translocation during the synthesis of proteins, cells affected by DT undergo apoptosis-driven cell death. DT is maximally expressed with low iron levels in the environment, and expression is supressed in the presence of high levels of iron [348]. It was believed that the gene *dtxR* was solely responsible for regulation of DT expression, but some recent publications have suggested that not only may toxin transcription have alternative regulation factors, but that *dtxR* may itself have other regulatory roles in additional to the regulation of iron metabolism and DT [348–350].

A recommended treatment for diphtheria is antibiotics, primarily penicillin or erythromycin, coupled with the rapid application of antitoxin for infections by toxigenic *C. diphtheriae* isolates [217]. In recent years, antitoxin has been in limited supply, especially across Europe, where reported diphtheria cases are much lower [217,218]. Multiple countries within Europe, including Spain and Belgium, have reported cases of diphtheria where antitoxin was not available in time for treatment [215,216]. These cases highlighted the vulnerability of children without full vaccine protection, and the delay in antitoxin administration was a major factor in both their deaths [215,216].

These cases came after multiple reviews highlighting gaps in antitoxin stockpiles across Europe due to the perceived low risk of disease occurrence. This was despite the fact that cases imported from countries where the disease is endemic had occurred sporadically across the region over previous years [217,218]. In addition, production of the antitoxin has dipped globally, as many pharmaceutical companies move away from the product due to low demand and profitability. The production process of diphtheria antitoxin has not progressed fundamentally from 1901, when Emil von Behring and Shibasaburō Kitasato's discovery of

animal sera offering anti-toxic effects in infected humans [212]. Antitoxin is still produced using equines today, making it expensive and time consuming to manufacture [214]. Alternative methods of producing antitoxin treatments, including mono- and polyclonal antibody-based therapies have been investigated, although none are currently in widespread production or use [214,219].

An assessment of genetic diversity within the *tox* gene in Corynebacterium at a global level could be valuable for assessing the potential for vaccine escape. Both the toxoid vaccine and antitoxin treatment are based on the toxin derived from a single allelic version of the gene, that carried by *C. diphtheriae* isolate PW8. By using our 502 strong *C. diphtheriae* genome sequence data set described in Chapter 3, we aimed to investigate the diversity of the *tox* gene using isolates collected across the globe. By using genomes from isolates collected across 122 years and 16 countries and territories, we can place gene variations into a spatial and temporal context. We also sought to investigate how the prevalence of non-toxigenic isolates had changed over time, as well as across countries and regions. By using our Indian collection from Chapter 3, we also aimed to focus in on the diversity of the diphtheria *tox* gene within a national context, within the country that reports the highest number of diphtheria cases per year.

Moving beyond gene diversity, we aimed to investigate what impacts the changes in the gene sequence might have on the amino acid sequence. To do so, we further investigated any SNPs that caused a non-synonymous amino acid change from the sequence of the PW8 vaccine strain *tox* gene. It is important to investigate the potential impact of any non-synonymous SNPs in the *tox* gene on the 3D crystal structure of the DT protein. This could have a very

real impact on the success and viability of current vaccine and anti-toxin formulations, as well as guiding future development of novel alternatives. Additionally, we investigated if the variation observed within the *tox* gene corresponded with differences in the integrated corynephage sequence of completed *C. diphtheriae* genomes.

4.2 Extracting the *tox* gene sequences from 502 *C. diphtheriae* isolates

The 502-genome collection described in Chapter 3, isolated across 122 years and 16 countries and territories, was used to investigate the diversity of the *C. diphtheriae tox* gene globally. In-silico polymerase chain reaction was used to extract the *tox* gene sequences using the in_silico_pcr script written by Simon Harris [106]. The primers were designed to encompass the entire gene and extract the full sequence, and were based on the *tox* gene annotated within NCTC 13129 [126]. The primers that were used are shown in Table 4.1.

Table 4.1: *Forward and reverse primers used to extract the C. diphtheriae tox gene*

Target	Forward Primer	Reverse Primer
tox	GTGAGCAGAAAACTGTTTGCGTCAA	TCAGCTTTTGATTTCAAAAAATAGC

Across the 502 genomes, *tox* genes were extracted from 291, 58% of the total collection. 211 isolates did not harbour the *tox* gene (*tox*$^{-}$). Among these *tox* gene positive (*tox*$^{+}$) isolates, 288

isolates contained a single copy of the gene. Three isolates contained two identical copies of the *tox* gene within their genomes. Two of these were isolated in Germany, while the other one was isolated in Switzerland. The 291 *tox* genes extracted were combined into a multifasta file, before being aligned using Clustal Omega [94]. Individual variants of the gene were identified and classified. After clustering the *tox* genes into variants, representatives from each group were aligned and SNPs called. After converting the *tox* gene nucleotide sequences to amino acid sequences using Seaview, SNPs identified within the variants were investigated individually, and any that caused a change to the amino acid sequence were labelled as non-synonymous SNP changes [351]. The graphical representation of this workflow is shown in Figure 4.1 below.

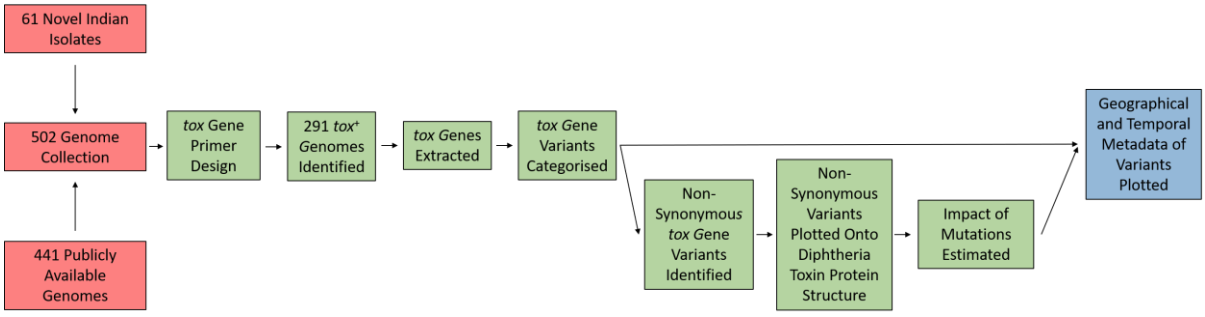

Figure 4.37*: Flow chart diagram for investigating the diversity of the C. diphtheriae tox gene present within the global collection of genomes. Red boxes represent the data collection phase, green the data analysis phase, and blue the data interpretation phase.*

4.3 *tox* gene diversity across the globe

18 variants of the *tox* gene were identified across the 291 *tox*$^+$ isolates. Group 16 (the variant carried by PW8) was the most common, being carried in 144 (~49.5%) of *tox*$^+$ isolates. The next highest respectively were Groups 11, 4, 15, 8, and 6. Of the three isolates that carried two identical *tox* genes, two of the genomes carried the Group 7 variant (Switzerland and Germany) while the other one from Germany carried a Group 16 *tox* gene. The number of isolates that carried each *tox* gene variant is shown in Figure 4.2. Of the 18 *tox* gene variant groups, six were found in only one isolate and 12 were found in less than 10 isolates. Collectively, these 12 variants were found across 31 genomes, making up ~11% of *tox*$^+$ genomes. While Group 16 was present in ~49.5% of *tox*$^+$ genomes, the other five most prevalent (4, 6, 8, 11, and 15) were found in 25, 13, 18, 37, and 23 genomes respectively, making up 116 (~40%) of the *tox*$^+$ isolates.

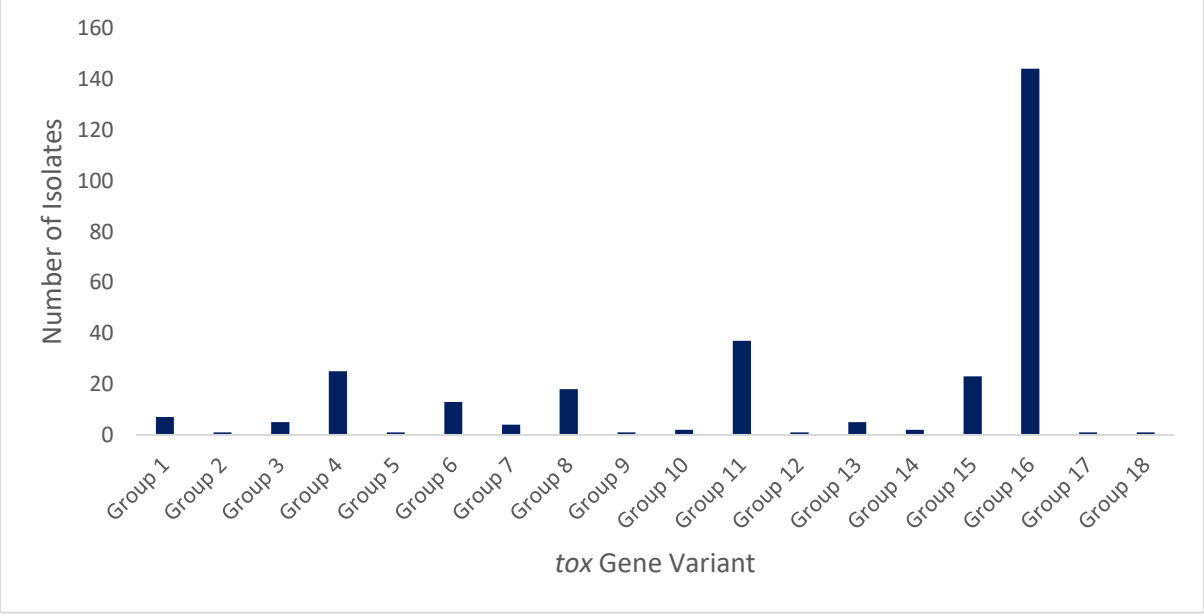

***Figure 4.38**: The number of isolates that carried each C. diphtheriae tox gene variant.*

Figure 4.3 shows the proportion of the 18 *tox* gene variants found across the 291 *tox*$^{+}$ genomes, along with the 211 *tox*$^{-}$ isolates, grouped per decade of isolation. The number of isolates per decade is also shown. Figure 4.4 shows the timeline of *tox*$^{+}$ and *tox*$^{-}$ genomes by their year of isolation. The diversity of *tox* gene variants detected increases significantly by decade by testing using a Pearson's product-moment correlation test (r (9) = 0.70, p = 0.02). There is however a much larger number of genomes derived from isolates in the more recent years, and as such this result may be due to the unavoidable sampling bias present in our dataset. Group 16 was the *tox* gene variant found in either the largest or joint-largest proportion of genomes per decade in all but the 1940s, 1960s, and 2000s, with the 1940s and 1960s being the only two decades where Group 16 was not observed. Both of these decades only had one *C. diphtheriae* isolate in our collection.

Of the 18, only three variants were not found within genomes isolated during the 2010s: Groups 1, 5, and 13. The first *tox*$^{-}$ isolate was found in the 1960s, before becoming more regularly present through the 1980s to the 2010s. There was an increase in the number of *tox*$^{-}$ genomes by decade by testing using a Pearson's product-moment correlation test (r (9) = 0.60, p = 0.05). While the highest proportion of *tox*$^{-}$ isolates (discounting the single 1960s genome) was in the 2000s (39 of the 69 isolates, 56.5%), the highest number of *tox*$^{-}$ isolates was found in the 2010s, at just under half of the 327 genomes from that decade (155 isolates, 47%). Despite this rise in *tox*$^{-}$ genomes being isolated, Group 16 remains the most predominant *tox* gene variant, although this may be due to a sampling bias towards India and Southeast Asia where the variant is much more commonly found among our collection.

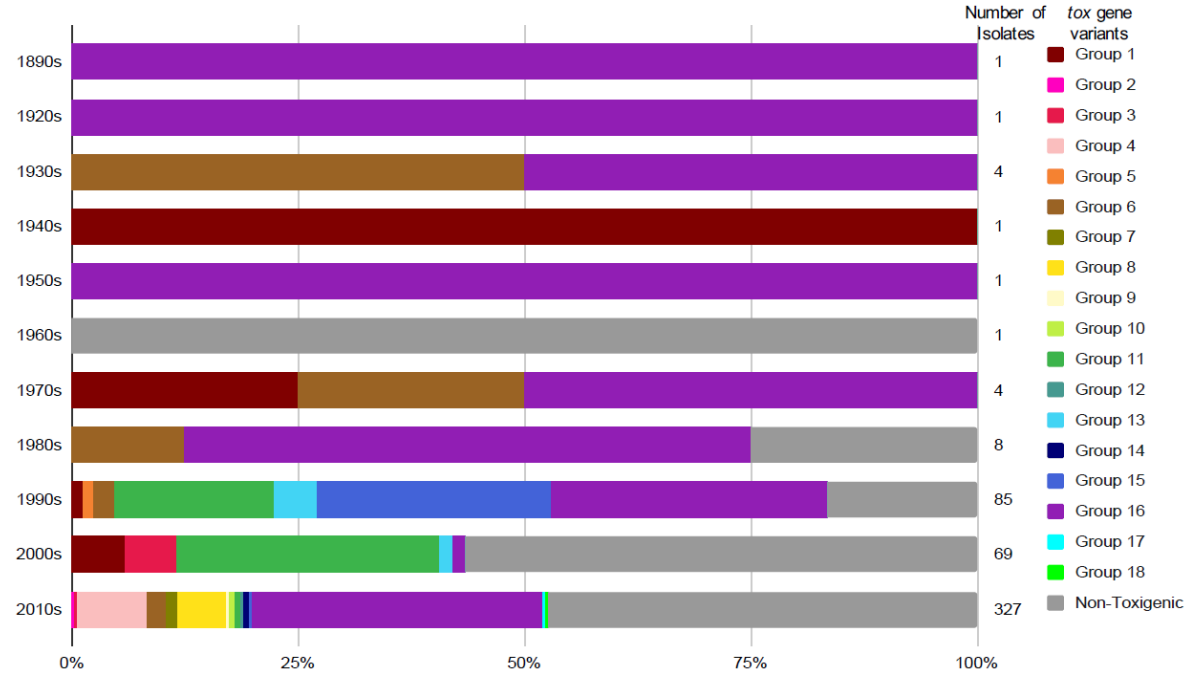

Figure 4.39: *The proportion of the 18 tox gene variants found across 291 tox⁺ and 211 tox⁻ isolates per decade, with the number of isolates per decade shown.*

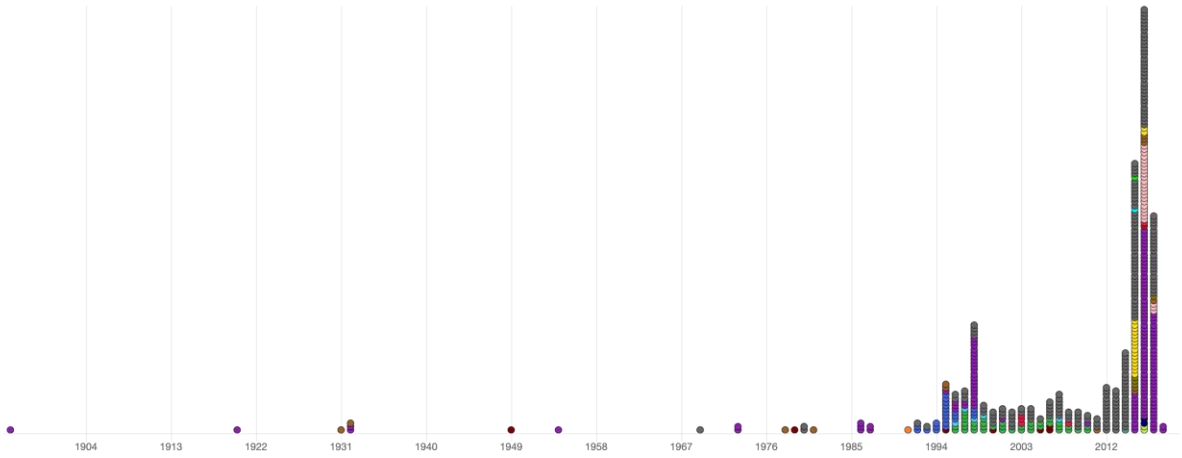

Figure 4.40*: Timeline of 502 C. diphtheriae genomes, coloured by the tox gene variant or tox⁻. Created using Microreact* [352].

Figure 4.5 show the 18 *tox* gene variants found across the 291 *tox*⁺ genomes, as well as the 211 *tox*⁻ isolates, by their country of origin. The markers are scaled by the number of isolates.

Some *tox* variants were only found within a single country, such as the Group 8 variant only found in 18 South African isolates. These isolates also all come from a single monophyletic group in Figure 3.5. Groups 17 and 18 are both only present in a single isolate each, both from Australia. This contrasts with the more widespread groups such as the Group 16 variant, which while being predominantly found across Asia (in India, Malaysia, and Vietnam), it was also found outside the region, harboured by a small number of genomes in the Northern Hemisphere: once in Belarus, twice in Germany, twice in the United Kingdom, and once in the United States of America. This pattern was also the case across the phylogeny in Figure 3.5, as these *tox* gene variants were found in multiple monophyletic groups and across multiple clusters throughout the tree. Much of the *tox* gene variation was in isolates from Asia, with 10 of the 18 groups contained within those genomes isolated from India, Malaysia and Vietnam. *tox*$^{-}$ genomes were much more commonly found in mainland Europe, Brazil, and Australia than among those genomes isolated in India and Southeast Asia. The difference between the number of non-toxigenic and toxigenic isolates in HICs and LMICs was found to be significant by testing using a chi-squared test (r (1) = 202.67, $p < 0.001$).

Figure 4.41*: Map of 502 C. diphtheriae genomes, coloured by the tox gene variant or tox⁻, and scaled by number of isolates. Created using Microreact* [352].

4.4 *tox* gene diversity within India

Only 4 *tox* gene variant groups were present within Indian *C. diphtheriae* in our collection. Figure 4.6 shows the breakdown of these variants by state, with the proportion of *tox* gene variants among the *tox*$^+$ genomes, as well as those that were *tox*$^-$. Among those isolates form Northern India, only two *tox* gene variants were found. Group 16 was the most prevalent, making up 58 of the 66 isolates, while Group 4 was the only other variant found, carried by eight of the isolates. All two of the Himachal Pradesh isolates, as well as all seven from Haryana, were Group 16, while 17 of the Delhi isolates and 32 of the Uttar Pradesh isolates also carried Group 16. Five isolates from Delhi and three from Uttar Pradesh carried the Group 4 *tox* gene variant. No *tox*$^-$ genomes were found among these isolates from the North. In the South of the country, four variants were found, as well as all of our Indian *tox*$^-$ isolates. In Tamil Nadu, 13 isolates carried the Group 16 *tox* gene variant, 11 carried the Group 4

variant, and one isolate was *tox*$^{-}$. In Kerala, 19 isolates carried the Group 16 *tox* gene variant, six carried the Group 4 variant, one the Group 2 variant, and two the Group 10 variant. Three isolates from Kerala were *tox*$^{-}$.

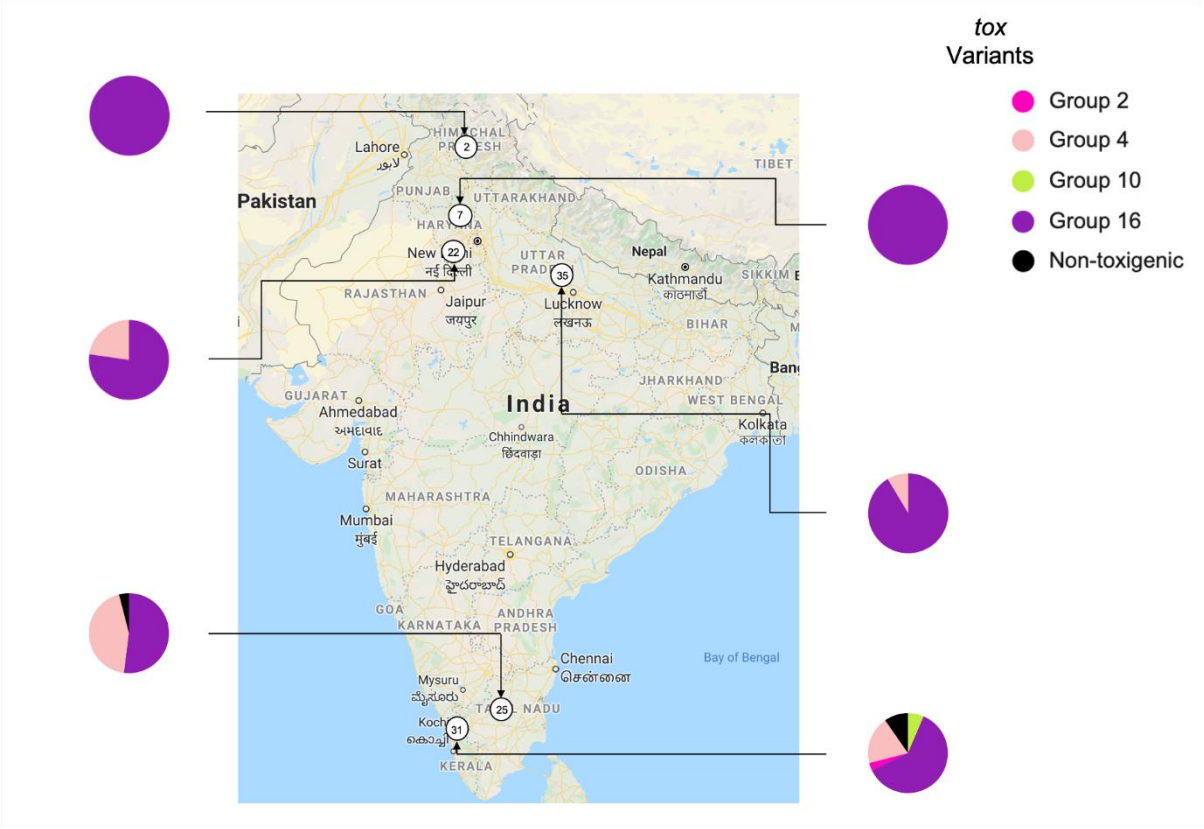

Figure 4.42*: The tox gene variant proportions represented in the 122 C. diphtheriae genomes from India at state level. The number of isolates from the six states sampled are shown in circles on the map. The map was taken from Google Maps* [341].

Across the 122 isolates, genomes carrying the Group 16 *tox* gene variant were isolated in all five years of isolation – 1973, 2015, 2016, 2017, and 2018. The three genomes that carried the Group 2 and Group 10 *tox* gene variants were isolated in 2016, while the Group 4 variants where isolated in 2016 (22 genomes) and 2017 (three genomes). The four *tox*$^{-}$ genomes were isolated in 2016. Figure 4.7 shows the timeline of these *tox* gene variant isolations.

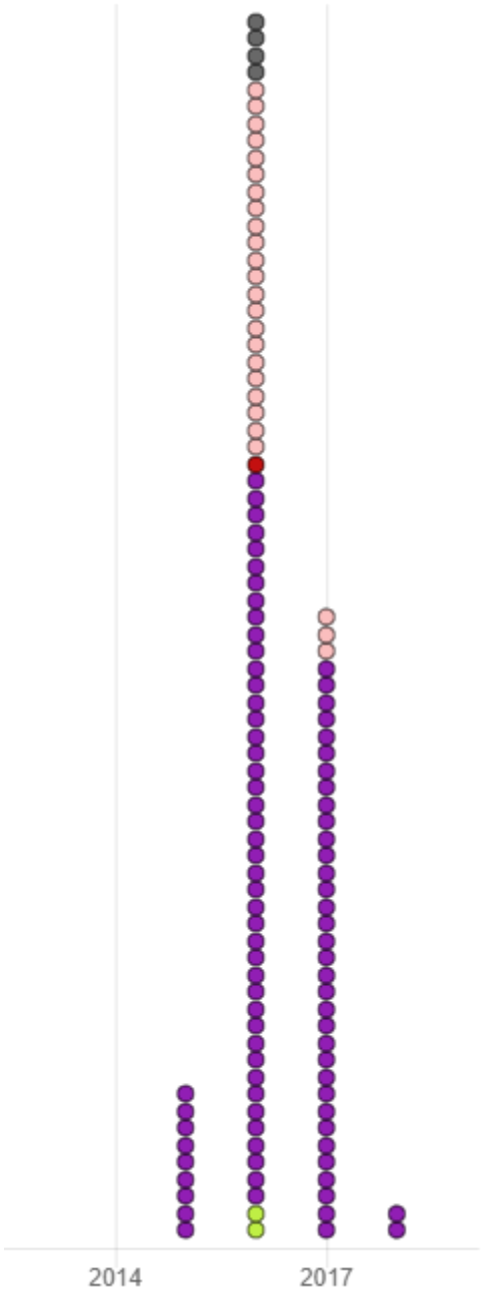

***Figure 4.43**: Timeline of 120 C. diphtheriae genomes, coloured by coloured by the tox gene variant or tox⁻. Two isolates from 1973 isolated in Himachal Pradesh were also included in the collection, which were both Group 16. Adapted from Microreact* [352].

4.5 Mapping the non-synonymous *tox* gene variants to the diphtheria toxin protein structure

To investigate the impact the variants of the *C. diphtheriae tox* gene may pose, we compared the amino acid sequence of each variant to Group 16. Any SNPs that lead to a change in the amino acid sequence were labelled as non-synonymous SNPs. The location of the change, the amino acid changed and which amino acid it was changed into was recorded.

Eight of the 18 *tox* gene variants were found to encode non-synonymous SNP changes. These were Groups 5, 7, 8, 13, 14, 15, 17 and 18. Of these, Groups 5, 8, 14, 15, 17, and 18 all contained one amino acid substitution, while Group 7 contained two. One of the Group 7 substitutions was shared with Group 8. The *C. diphtheriae* carrying these non-synonymous variants were isolated between 1991 – 2017, and from seven countries: Australia, Belarus, Germany, Malaysia, South Africa, Switzerland, and Vietnam. In total, 55 genomes were positive for these non-synonymous *tox* gene variants. Both Groups 5 and 15 were among the six most common *tox* gene variants globally. Figures 4.8 and 4.9 show the map and timeline of isolation for these non-synonymous *tox* gene variant-carrying genomes. All the non-synonymous Group variants were isolated from within an individual country except Group 7, which was isolated in three genomes from Switzerland in 2015, as well as in one genome from Germany in 2017. Malaysia and Australia were the only two countries to have multiple non-synonymous *tox* gene variants. In Australia, one genome carried the Group 17 variant, and one genome carried the Group 18 variant, both isolated in 2015. In Malaysia meanwhile, one isolate from 1991 carried the Group 5 variant, and two isolates from 2016 carried the Group 14 variant. Among the others, Group 8 was only carried by 18 genomes from South Africa (16 from 2015, two from 2016), and Group 13 was only carried by five genomes

isolated from Belarus (one each from 1996, 1997, 1998, 1999, and 2007). Group 15 was carried by the largest number of isolates among the non-synonymous *tox* gene variants, found in 23 genomes isolated in Vietnam. Two were isolated in 1992, one in 1993, three in 1994, 10 in 1995, 3 in 1996, and the last 3 in 1998, while for one bacterium the year of isolation was unknown.

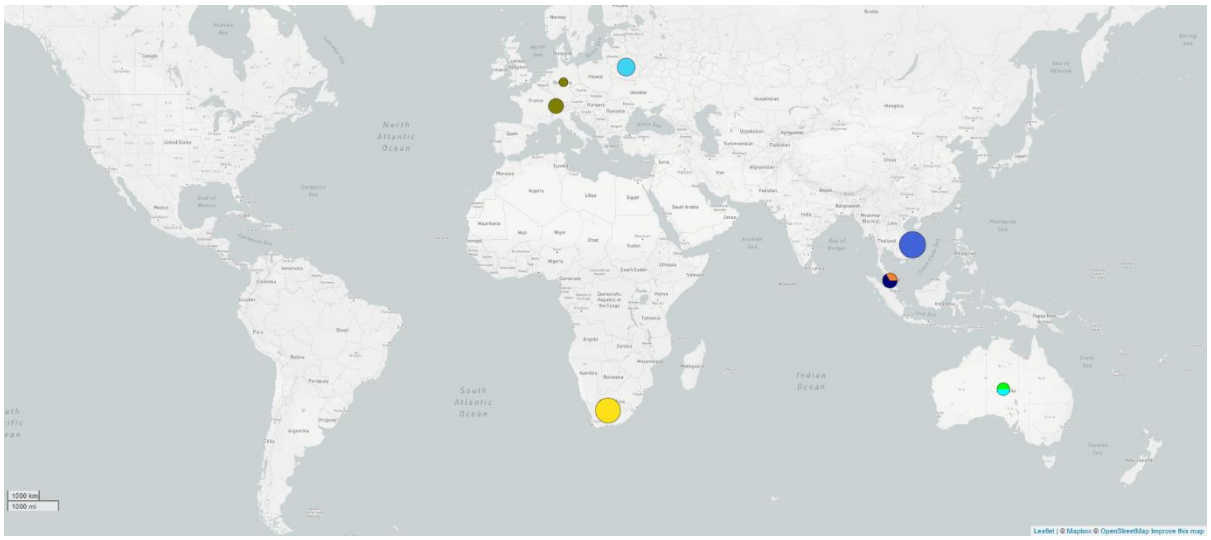

***Figure 4.44**: Map of 55 C. diphtheriae genomes carrying non-synonymous tox gene variants, coloured by the tox gene variant, and scaled by number of isolates. Created using Microreact* [352].

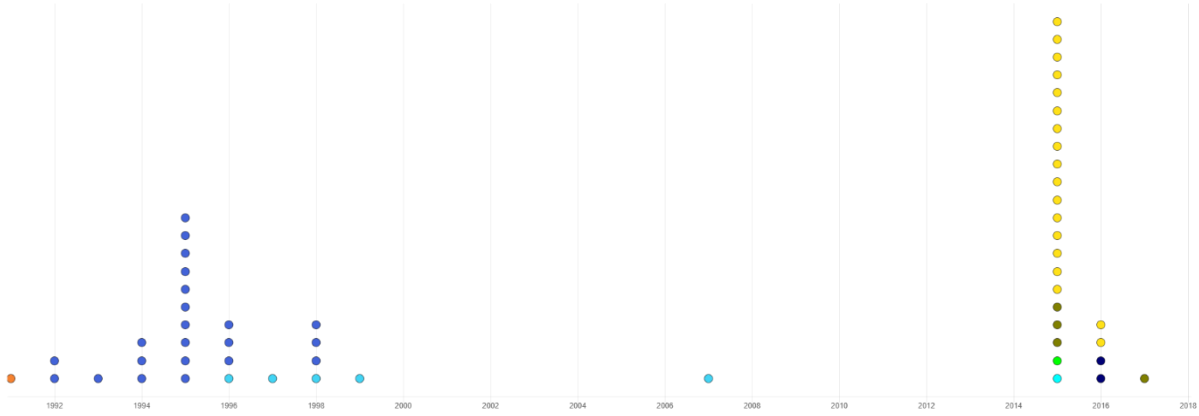

***Figure 4.45**: Timeline of 55 C. diphtheriae genomes carrying non-synonymous tox gene variants, coloured by the tox gene variant. Created using Microreact* [352].

After identifying the non-synonymous variants and the site of their mutations, we mapped the sequences to the diphtheria toxin protein to assess the likely impact of those mutations. Of the eight non-synonymous *tox* gene variants, two Groups had amino acid changes that occurred within the signal sequence of the gene, and as such could not be mapped to the protein structure. Group 13 had a base pair deletion that lead to a defunct *tox* gene in the five Belarussian isolates, previously reported by Grosse-Koch *et al* as resulting in the isolates being NTTB [130]. Group 14 meanwhile had a histidine (H) that had changed to a tyrosine (Y). The impact of this mutation could not be assessed using the tools available.

Figure 4.10 shows the six remaining non-synonymous *tox* gene variants mapped to the *tox* gene protein structure, using the Protein Data Bank (PDB) model 1XDT and UCSF ChimeraX [331–333]. After mapping, we used PHYRE2 and SuSPect to estimate the impact of these mutations on the protein structure at each individual site [334,335]. The shared amino acid mutation contained in Group 7 and 8 was estimated to have a low impact on the protein structure. The second Group 7 mutation, as well as those carried by Groups 5, 15, and 18

were all estimated to have a moderate level of impact. The Group 17 mutation was the only amino acid substitution within our eight non-synonymous *tox* gene variants to be estimated as having a high impact on the structure of the toxin protein. It is notable though that the location of the mutations in Groups 15, 17 and 18 had a much higher average impact of mutation posed by other amino acid substitutions. Many of these other potential substitutions showed a high risk of the mutation impacting the protein structure at this site.

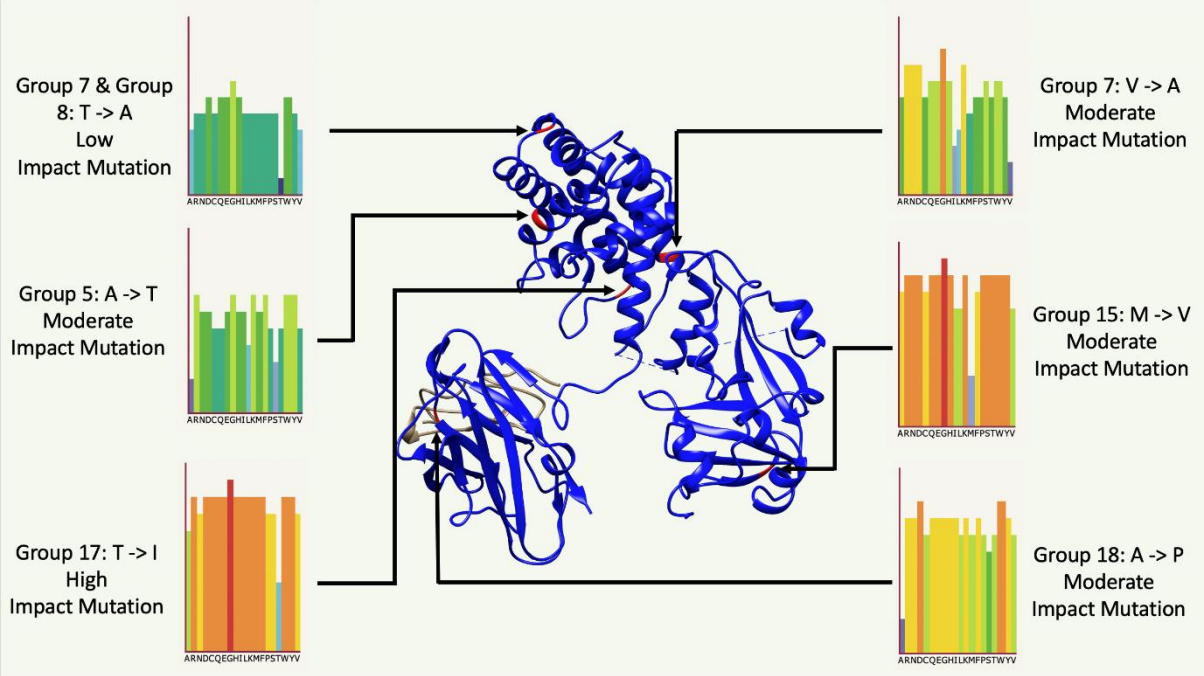

***Figure 4.46**: Six non-synonymous tox gene variant mutations plotted onto the diphtheria toxin model 1XDT (https://www.rcsb.org/structure/1xdt) from the Protein Data Bank using PHYRE2* [331,334]*. The impact of these mutations is estimated by SuSPect, with a gradient per mutation of low (dark blue) to high (orange/red)* [335]*.*

4.6 Variation within the corynephage

The *tox* gene is carried by corynephages that can integrate into the genome of infected *C. diphtheriae* bacteria. To investigate if the diversity present in the *tox* gene corresponded with

diversity in the integrated corynephage, we mapped 11 publicly available complete *C. diphtheriae* genomes that carried the *tox* gene from NCBI Genbank to the corynephage sequence annotated in C. diphtheriae isolate NCTC 13129, using BWA [126,353,354]. By using complete genomes, we can have the highest reliability possible that the extracted mapped sequences are accurate. The 36,570 nucleotide base long alignment was used to create a maximum likelihood phylogeny over 1000 pseudo-bootstrap replicates using IQ-TREE, before the country and decade of isolation, as well as which *tox* gene variant was carried, were annotated in iTOL [102,112].

Figure 4.11 shows the maximum likelihood phylogeny of these 11 complete tox^+ genomes. There were no major phylogenetic correlations between the decade of isolation or the country of isolation. Corynephages from large temporal and spatial distances clustered together. Additionally, the *tox* gene carried by these corynephages does not correspond to the phylogenetic structure, with the more common variants among this 11 genome collection, Groups 16 and 6, found in distantly related isolates across the tree.

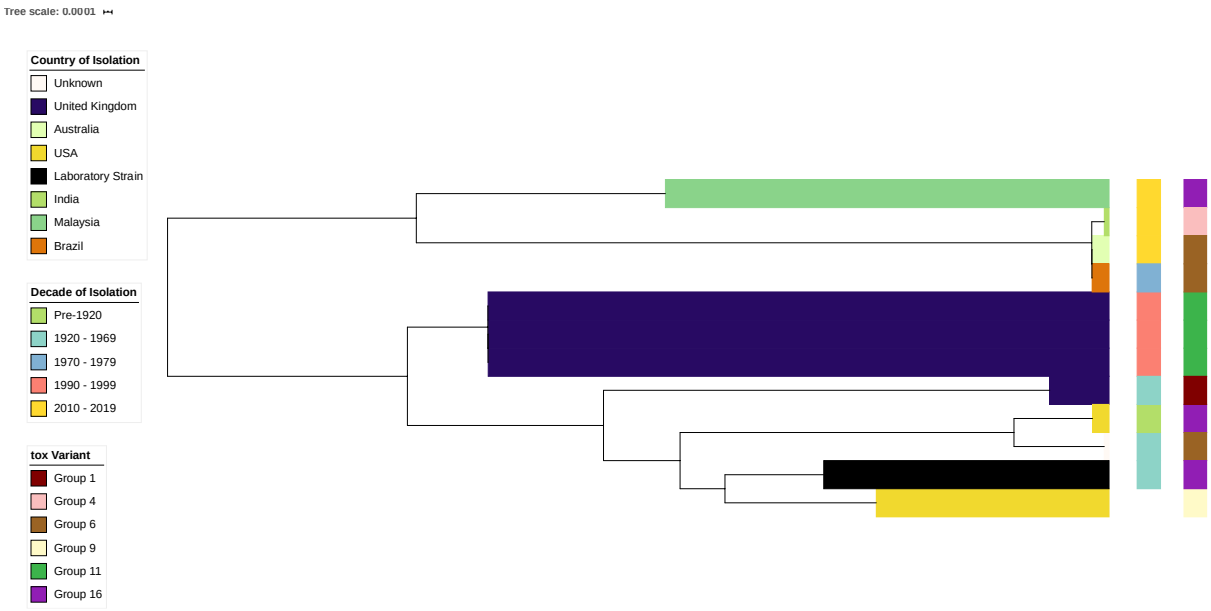

***Figure 4.47**: Maximum likelihood phylogeny of corynephage from 11 toxigenic completed genomes, showing the country and decade of isolation along with the tox gene variant carried. Created using IQ-TREE over 1000 pseudo-bootstrap replicated and annotated in iTOL [102,112]. The scale bar shows substitutions per site.*

4.7 Discussion

By using the 502 *C. diphtheriae* genome collection described in Chapter 3, we were able to investigate diversity within the *tox* gene, as well as factoring in those of genomes without it, from across 16 countries and territories across the globe, covering 122 years of isolation. The 291 genomes positive for the *tox* gene made up 58% of the total collection, with the other 211 isolates being *tox*$^{-}$. We were able to identify 18 variants of the *tox* gene, with Group 16 (which is carried by the strain PW8 used in vaccine production) being the most common within our collection, carried by 49.5% of *tox*$^{+}$ isolates. While this is only just under a majority, it does show that there are a large number of *tox* gene variants being carried in the world that are not exact matches to that which is used in the vaccine production process. The

five next-most prevalent *tox* gene variants (4, 6, 8, 11, and 15) being found in ~40% shows that while there is a high number of variants, many are found in a small number of isolates. The fact that some *tox* gene variants were country specific, such as Group 8 in 18 South African isolates, or the two isolates from Australia that carried Groups 17 and 18, possibly suggests that these variants have not yet been circulated internationally, although further analysis of neighbouring countries will be required to establish this. Other groups, such as 16 and 4, have spread across the world and can be found across multiple distantly related monophyletic groups (Figure 3.5). This data reinforces the observation that human transmission has spread *C. diphtheriae* successfully across the globe and has resulted in numerous clones co-existing in a similar location at the same time, while remaining genetically distant.

The diversity of our *tox* genes increased significantly by decade, and while this may be due to the temporal sampling bias present within the collection, it is important to assess how this continues into the future. The number of *tox*$^{-}$ genomes also increased by decade, suggesting that there may be selective pressures being applied to populations of *C. diphtheriae*, both for *tox* gene mutations as well as promoting the rise of entirely non-toxigenic isolates to the fore in some regions of the world. The difference between the number of non-toxigenic and toxigenic isolates in HICs and LMICs being significantly different adds another piece of evidence to the selection pressure existing, as HICs have historically shown a lower number of diphtheria cases, often attributed to higher levels of vaccine coverage. If *C. diphtheriae* is to re-emerge in these areas, it will have to do so in a differently adapted form, as is the case in the rise of non-toxigenic case numbers reported. In Europe, *tox*$^{-}$ genomes were much more common, and this is true also true in Australia as well as in the Americas. These locations also reported a much lower number and proportion of the Group 16 *tox* gene variant.

Countries in Asia meanwhile have shown the highest numbers of *tox* gene variants present compared to other regions, corresponding with the highest numbers of diphtheria cases reported annually. As most of the world now has the diphtheria toxoid vaccine as part of national immunisation programs, and vaccine coverage levels globally have been increasing, this selection pressure on the diphtheria toxin is bound to increase even more in the future.

Within India, Group 16 remains the dominant *tox* variant, and non-toxigenic cases are much rarer among the public genome record. Among the genomes from Northern India, only Group 4 and Group 16 were present, with no other variants or *tox*$^{-}$ genomes present. This contrasts with the South of the county, where Groups 2 and 10, as well as four *tox*$^{-}$ genomes, were all found in 2016. While these other variants only made up a small number, it suggests that in recent years *tox* gene variants have been mutating or introduced from external sources, along with the appearance of non-toxigenic isolates causing disease, suggesting that selection pressures may be becoming more common.

Among our non-synonymous isolates, the importance of further studies is clear. While only one isolate carried Group 17, which was the only variant estimated to have a mutation that could have a high level of impact on the protein structure, four other groups were estimated as having moderate impact mutations, while only one was estimated as having a low impact. While these results do not seem to highlight any concern in the efficacy of the currently used Group 16 diphtheria toxoid vaccine formulation, it is vital that further in vitro and in vivo studies are carried out to assess the real-world impact these mutations have on the diphtheria toxin protein structure. This is imperative, as the impact that any protein structural change

may have on the vaccine or antitoxin treatment needs to be understood before it becomes a wide-scale national and global health problem.

The rise in both *tox* gene diversity and *tox*$^-$ genomes do forecast a potential future where the antitoxin treatment and vaccine become less effective, or indeed are fully escaped. As more genomes of *C. diphtheriae* are sequenced, additional variants may be revealed, and these must also be investigated to assess how much of an impact they have on the final protein structure. While there is a temporal bias in the collection, the number of *tox* gene variants was shown to be increasing over time, and new variants could theoretically mutate and become more widely distributed at any time. As the coverage of the diphtheria toxoid vaccine continues to rise globally, there will be an ever-increasing selection pressure towards a diphtheria toxin that can evade the vaccine, as well as the antitoxin treatment. This pressure will also potentially be encouraging the evolution of non-toxigenic *C. diphtheriae* isolates, and further detailed investigations into the mechanisms of how non-toxigenic *C. diphtheriae* causes infection and symptoms are needed. While the diversity in the *tox* gene does not seem to be tied to diversity in the corynephage, the low number of complete genomes available make this conclusion preliminary. As mentioned above, the further sequencing of *C. diphtheriae* genomes, as well as the ever-improving technologies involved, will allow future analysis to dive deeper into any links between the *tox* gene evolution and its original viral host.

These results and conclusions highlight how important continued surveillance and analysis is into the diphtheria toxin. While much has been known about the agent for many decades physiologically, it is imperative that we remain vigilant into current and future evolution.

These conclusions also highlight the important of a plan B in hybrid toxin-based vaccine candidate selection for any future refinement of the toxoid formula, as well as alternative and more readily available anti-toxin treatments. Previous research has highlighted the need for better passive protection approaches, and these must be pre-emptively prepared for action. This is in case any non-Group 16 non-synonymous variant of the *tox* gene that allows for treatment and vaccine escape ever becomes more prevalent.

5.0 O1 *Vibrio cholerae* in Ghana

5.1 Introduction

Though cholera has been effectively eradicated in many HICs, the disease remains persistent in many LMICs across the world. Caused by the bacterium *V. cholerae* that is transmitted through contaminated food and water, the severe dehydration brough about by diarrhoea (often termed 'rice water stool', a characteristic symptom of cholerae) can prove fatal without the proper administration of fluid and electrolyte replacement therapies [157,355]. The incubation period of *V. cholerae* infection can range from between 12 hours and 5 days, and the bacteria can remain stable within aquatic environments for large periods of time, persisting in reservoirs including lakes, rivers, and stagnant water [178,356]. Climate change is believed to be playing a major role in the spread of *V. cholerae* bacteria, as coastal and other aquatic environments become more favourable to the bacteria with rising sea levels and ever-more changeable weather systems [357]. Cholera is a disease heavily linked to poverty,

Of the over 200 serogroups identified across the species, only O1 and O139 *V. cholerae* have been shown to cause epidemics [149]. Of these two, almost all outbreaks are due to the traditional O1 serogroup, and it is believed to be the cause of all seven pandemic waves of cholerae identified throughout history [149]. The current seventh pandemic began in 1961, and genomic analysis indicates three waves of independent but overlapping transmission events have spread *V. cholerae* across the world [120,358]. At least 12 transmission events have been linked to introductions into Africa, and cholera is likely endemic in numerous regions across the continent [120,124]. Indeed, the World Health Organisation (WHO) reported that in Africa

between 1970 – 2012, there were around 3.7 million cases, leading to 155,000 deaths [359]. These outbreaks especially affected coastal and river-side cities, as cholera spread along water channels and trade routes before being transmitted into more inland, land-locked settlements [124,178,360,361].

Ghana is a key exemplar of these circumstances in West Africa. From 1979 – 2015 the WHO published figures show 167,000 cholera cases, resulting in 5,851 deaths [359]. An interesting phenomenon that occurs in the country however is sustained periods with complete cholera absence reported. A four-year absence occurred between 1986 – 1989, before a one-year absence occurred again in 2013. During 2013, 157 cases of suspected cholera were investigated and analysed at the National Public Health Reference Laboratory in Ghana (NPHRL), but they were all confirmed to be not *V. cholerae* by laboratory analysis. Despite this, less than 12 months later in 2014, 50,000 cases of cholera were reported across nine of the 10 regions in Ghana, in one of the largest single outbreaks in the country's history. While the death rate was lower than previous outbreaks, it took until the start of November 2015 for the outbreak to be declared officially over [362]. This large-scale outbreak involved cholera cases 2.3 times above the national average, while the months after the four-year gap in 1986 – 1989 reported cases numbered 1.4 times the national average. These highlight the importance of continued and constant preparation and sampling to aid in cholera control and prevention.

Despite Ghana's history of cholera outbreaks, studies investigating the dynamics of epidemic cholera both in the country and in relation to the ongoing global seventh pandemic have been limited [361,363,364]. Despite the growing availability of whole genome sequencing technologies, *V. cholerae* sequences from Ghana are incredibly rare. Weill *et al*'s large-scale continental

study of the 'genomic history of the seventh pandemic of cholera in Africa' included four *V. cholerae* genomes from Ghana, two isolated in 1970, and two isolated in 1971. With Ghana's mix of coastal, river-side, and landlocked cities, coupled with the curious gaps where no cholera cases are observed shortly before large scale outbreaks occur presents an important potentially missing piece in the puzzle of understanding cholera across Africa.

Recent studies of *V. cholerae* outbreaks in other areas of the world have also highlighted the increasing development of antimicrobial resistance in *V. cholerae*, something that has not previously been investigated widely in Ghana [123,124]. By utilising the availability of molecular and computational analysis tools, we investigated how 127 novel *V. cholerae* isolates taken from outbreak cases in Ghana fit alongside 509 global representatives of the three waves of the 7th Pandemic (including the four previous isolates from Ghana) and identified the development of antimicrobial resistances within Ghanaian isolates over time. We also investigated the cholera toxin gene variants carried by our Ghanaian isolates, as well as using BEAST analysis to estimate introduction dates for these Ghanaian genomes.

5.2 Creating a representative collection of *V. cholerae*

127 *V. cholerae* were randomly selected from those isolated in Ghana between 2010 – 2016, including 124 from clinical cases dating between 2010 – 2015, and three from environmental sampling undertaken in 2016. DNA from these isolates were sequenced at the Wellcome Sanger Institute using an Illumina HiSeq v4 platform producing short read data. A literature review of major *V. cholerae* studies from across the world was undertaken, and appropriate genomic sequences were extracted from the publicly available databases National Center for

Biotechnology (NCBI) Genbank and European Bioinformatics Institute (EMBL-EBI) European Nucleotide Archive, including the four previously published Ghanaian isolates from 1970 and 1971 [288,318].

The metadata of these 509 isolates were acquired and recorded, before these genomes were combined with our novel Ghanaian genomes in a 636 strong globally representative collection. The oldest genome was from a *V. cholerae* isolated in 1957 and the most recent in 2016, giving a span of 59 years across 55 countries.

Most of the *V. cholerae* within our collection were isolated during more recent decades. One genome was from a *V. cholerae* isolated during the 1950s, 13 during the 1960s, 45 during the 1970s, 42 during the 1980s, 118 during the 1990s, 153 during the 2000s, and 252 during the 2010s. Geographically, our genomes represented *V. cholerae* isolated across five continents, incorporating 36 African countries, nine Asian countries, one European country, two North American countries and six South American countries represented among the collection. As we aimed to understand how our novel Ghanaian isolates fit within a global context, it was important to have a large number of African countries represented to give a more local context as well.

By bringing together all these genomes from numerous important *V. cholerae* studies published from across the globe, we can begin to determine how our novel Ghanaian isolates fit alongside those from neighbouring countries and beyond, as well as where they fit within the three waves of the ongoing 7th Pandemic.

5.3 Ghanaian *V. cholerae* in a global context

Once our 636-genome collection had been assembled, we began constructing our *V. cholerae* phylogeny. Utilising the widely used methodology of mapping genomic sequences to a reference genome for the construction of our genome alignment, we used the N16961 *V. cholerae* isolate as our reference. We used SMALT mapping, running the 'multiple_mapping_to_bam' script developed by Simon Harris at the Wellcome Sanger Institute, before running the 'join_dna_with_indels' script to produce a 636-strong multi-genome alignment [323,324]. To mask the impact that recombination would have on our phylogeny construction, we used Gubbins to remove areas of recombination from our alignment [321]. The subsequent SNP output from Gubbins was an alignment 4,676 bases long, and this was used as the foundation of our phylogeny.

Taking this SNP alignment, a phylogeny was created using IQ-TREE over 1,000 pseudo-bootstraps. Simultaneously, *in silico* gene presence and absence testing for cholera toxin variants and AMR determinant genes took place. This workflow is depicted graphically in Figure 5.1.

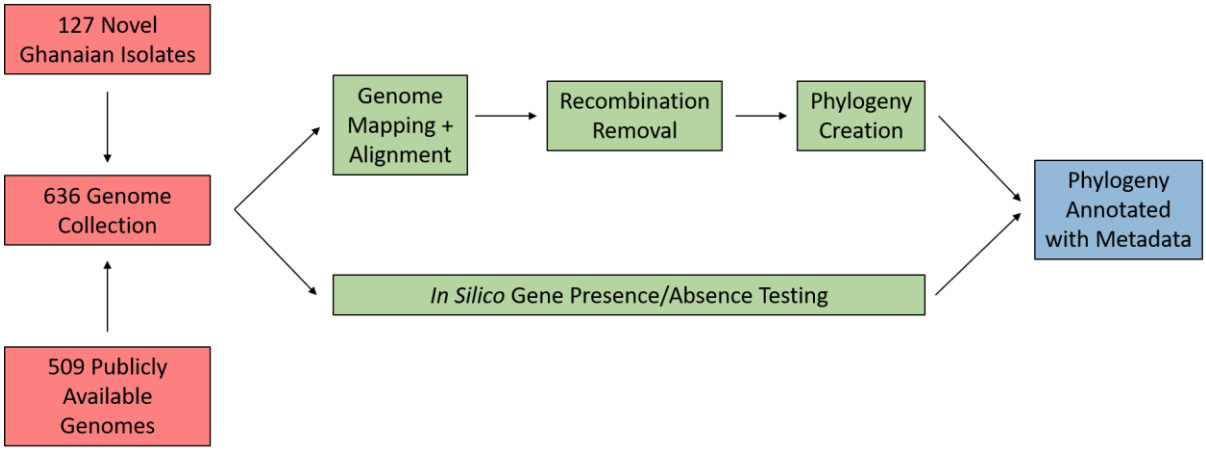

***Figure 5.48**: Flow chart diagram for generating the Ghanaian V. cholerae isolates and the globally representative phylogeny, as well as investigating the presence and absence of key genes. Red boxes represent the data collection phase, green the data analysis phase, and blue the data interpretation phase.*

Figure 5.2 shows the SNP phylogenetic tree, after recombination spots has been removed, of 636 *V. cholerae* genomes, annotated along with the waves of the 7th Pandemic, as well as the region the genomes were isolated from. Ghanaian isolates (including the four from 1970 and 1971) are labelled as 'this study' in the bolder orange colour. While the 1970s Ghanaian *V. cholerae* isolates clusters in Wave 1 with other older isolates, the modern Ghanaian *V. cholerae* isolates clustered into three distinct Clades (Clades 1, 2, and 3) within Wave 3. These can be seen in Figure 5.2 as dashed line boxes.

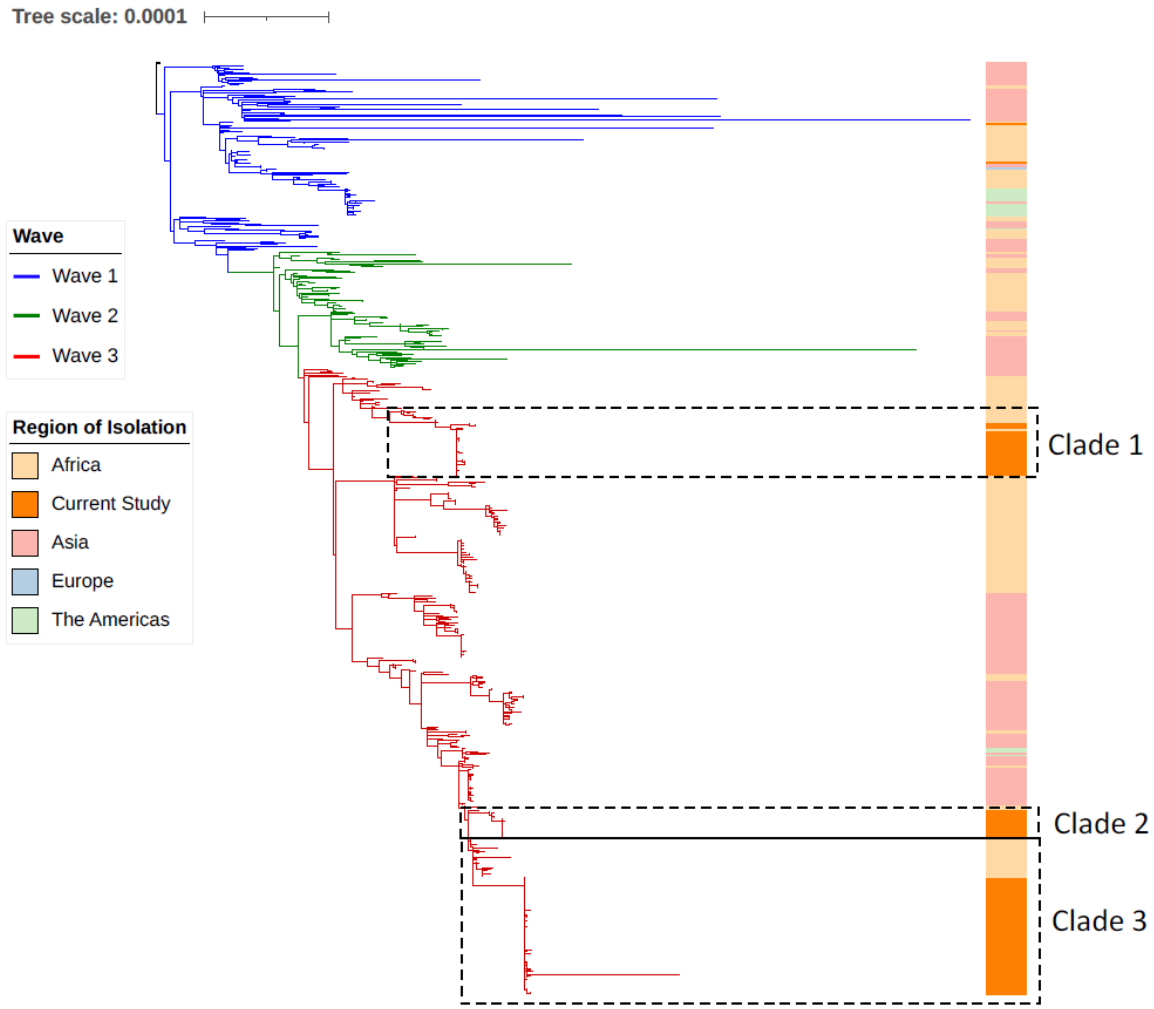

***Figure 5.49**: The maximum likelihood phylogenetic tree based on the mapped recombination-free SNPs from the 636 global V. cholerae genome collection. Branches are coloured by the wave of the 7th Pandemic, and the phylogeny is annotated with the region of isolation for each genome. Ghanaian isolates (including the four from 1970 and 1971) are labelled as 'this study'. The three Clades of 2010 – 2016 Ghanaian V. cholerae isolates are shown boxed. The scale bar shows substitutions per site.*

To determine the population clustering structure of the phylogeny in Figure 5.2, BAPs analysis was undertaken, and the results can be found in Figure 5.3 [337,365]. BAPs analysis identified 7 clusters, with Clade 1 as part of cluster 7, and Clades 2 and 3 as part of Cluster 4. Clade 1 contained genomes reported as part of the introduction event 9 (T9) previously, while Clade 3 contains isolates previously identified as being part of introduction event 12 (T12).

Based on the this, Clade 2 can similarly be suggested as being part of T12 [124].

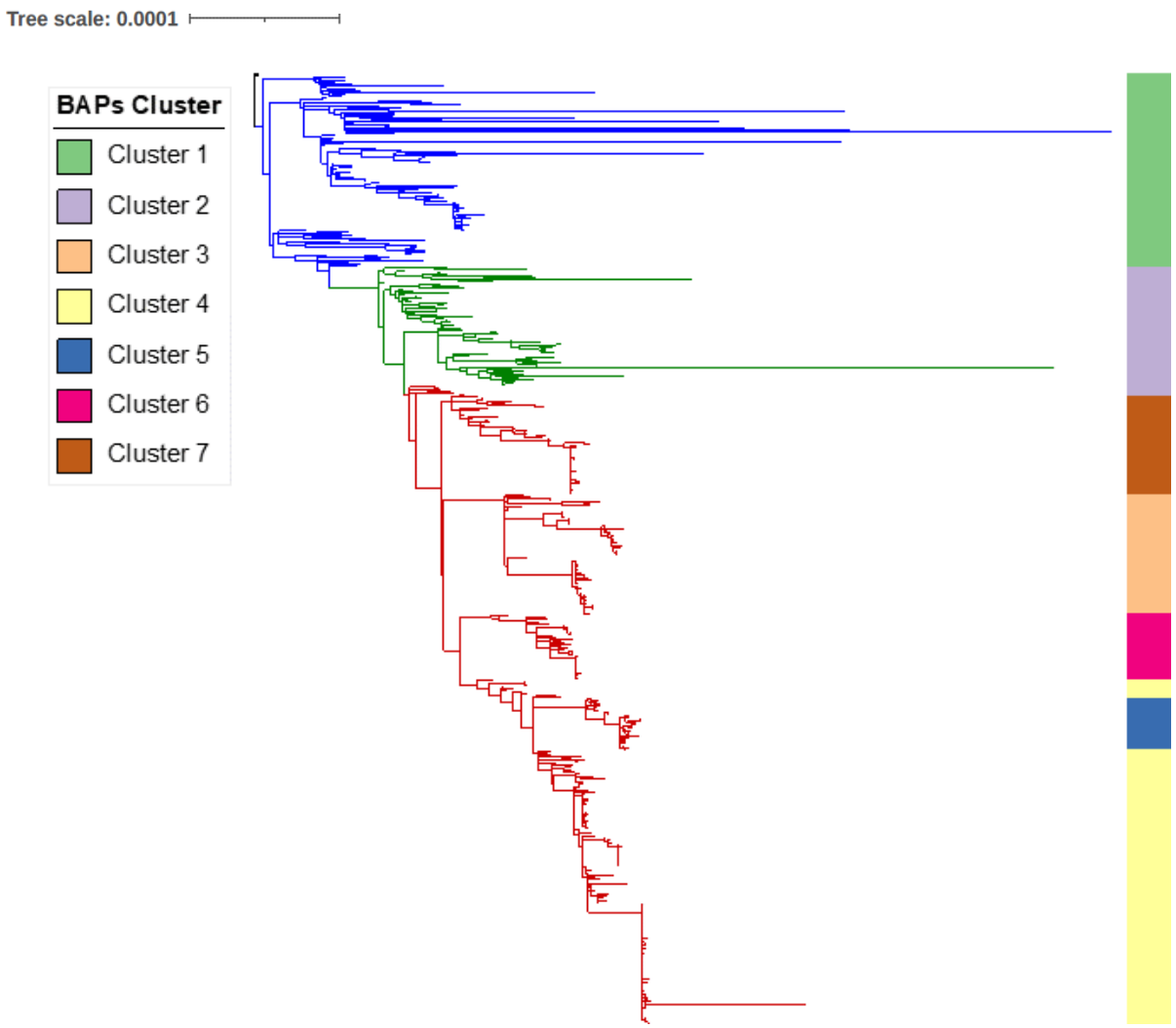

***Figure 5.50**: BAPs clusters plotted alongside the inferred global phylogeny of 636 V. cholerae isolates shown in Figure 5.2. Seven clusters were identified by BAPs. The scale bar shows substitutions per site.*

Figure 5.4 shows Clades 1, 2, and 3 in more detail, with the country and year of isolations annotated upon the snipped trees. The closest relatives of the Ghanaian isolates were from Benin, Cameroon, Niger, Nigeria, and Togo. *V. cholerae* isolated in Ghana during the 2010 – 2012 outbreak all clustered into eighter Clade 1 or Clade 2, while isolates from 2014 and

2015 were all contained within Clade 3. Interestingly, the three *V. cholerae* isolated from the environment in Ghana (marked by red stars on Figure 5.4) clustered within Clade 2, alongside those from the 2010 – 2012 outbreak. Outside of Ghana, other *V. cholerae* genomes inside Clade 1 were isolates as early as 2001 and 2005, and included isolates from Chad, Cameroon, Equatorial Guinea, the Republic of the Congo and Nigeria. Clade 2 meanwhile contained isolates from 2011 and 2012 from Togo and Chad, while Clade 3 clustered alongside isolates from 2010 and 2011, from Cameroon, Niger, and Togo.

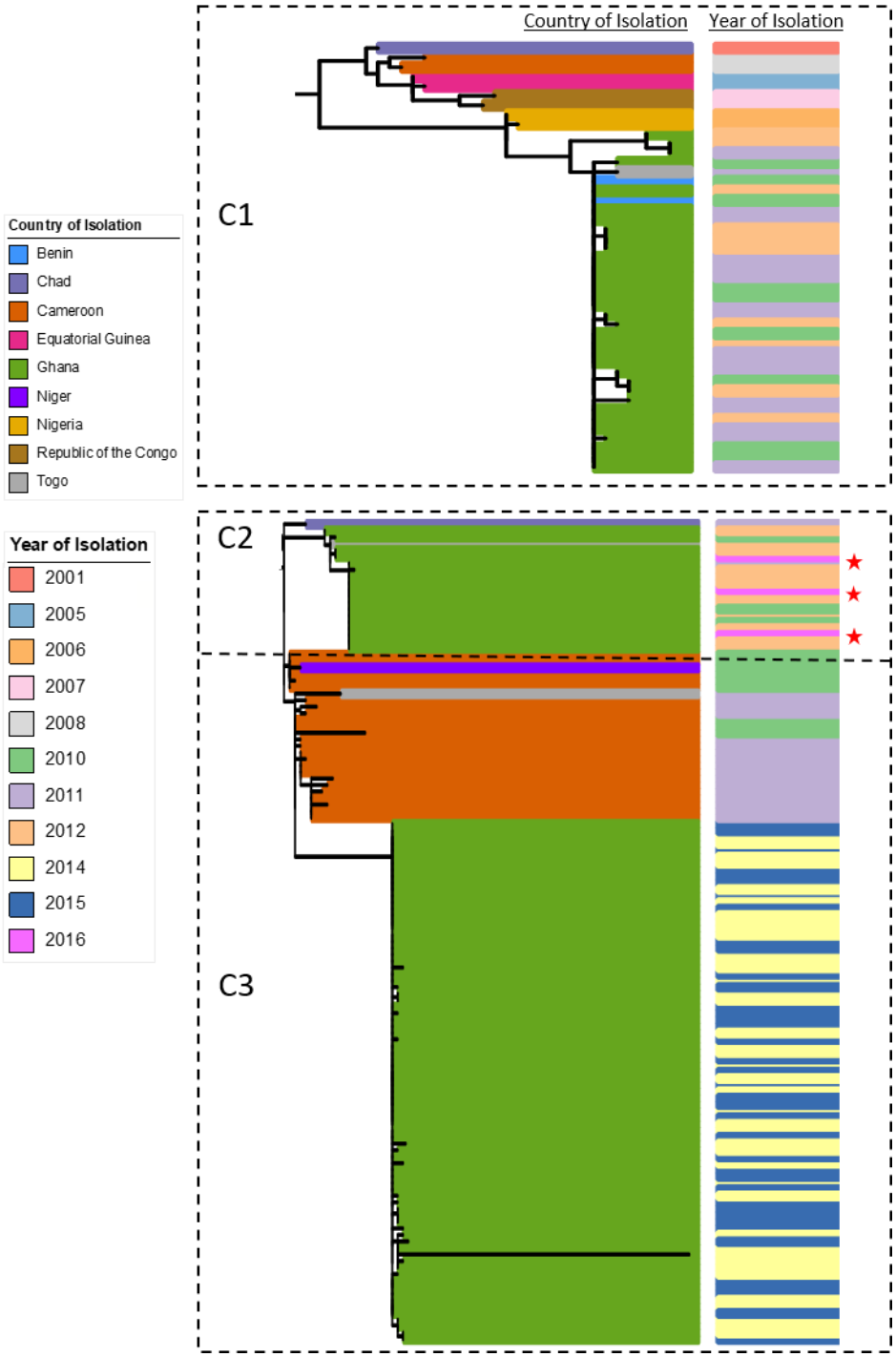

***Figure 5.51**: Subtrees of Ghanaian V. cholerae Clades 1, 2 and 3, trimmed from the maximum likelihood phylogeny in Figure 5.2. The country of isolation and year of isolation are shown annotated. Red stars designate the three environmental isolates from Ghana, isolated in 2016.*

5.4 Time scaled phylogenetic analysis

Following on from the creation of our maximum likelihood phylogeny in Figure 5.2, we undertook a time scaled phylogenetic analysis using BEAST on the whole phylogeny to estimate the introduction timing of our three Clades. The evolutionary rate across the whole phylogeny was estimated to be 6.376×10^{-7} substitution per site per year. Figure 5.5 shows the 95% confidence interval estimate of the most common ancestors of Clade 1 and Clades 2 and 3, as well as the estimated ranges of each Clade's introduction into Ghana. The most recent common ancestor of Clade 1 was estimated to have existed between February 1996 and August 2000, and the descendants were estimated to have been introduced into Ghana between January 2007 and April 2009. For Clades 2 and 3, the last common ancestor of both Clades was estimated to have existed between June 2005 and March 2008, while Clade 2 was estimated to have been introduced into Ghana between 2009 and December 2009. Clade 3 meanwhile was estimated to have been introduced to the country between July 2011 and May 2013.

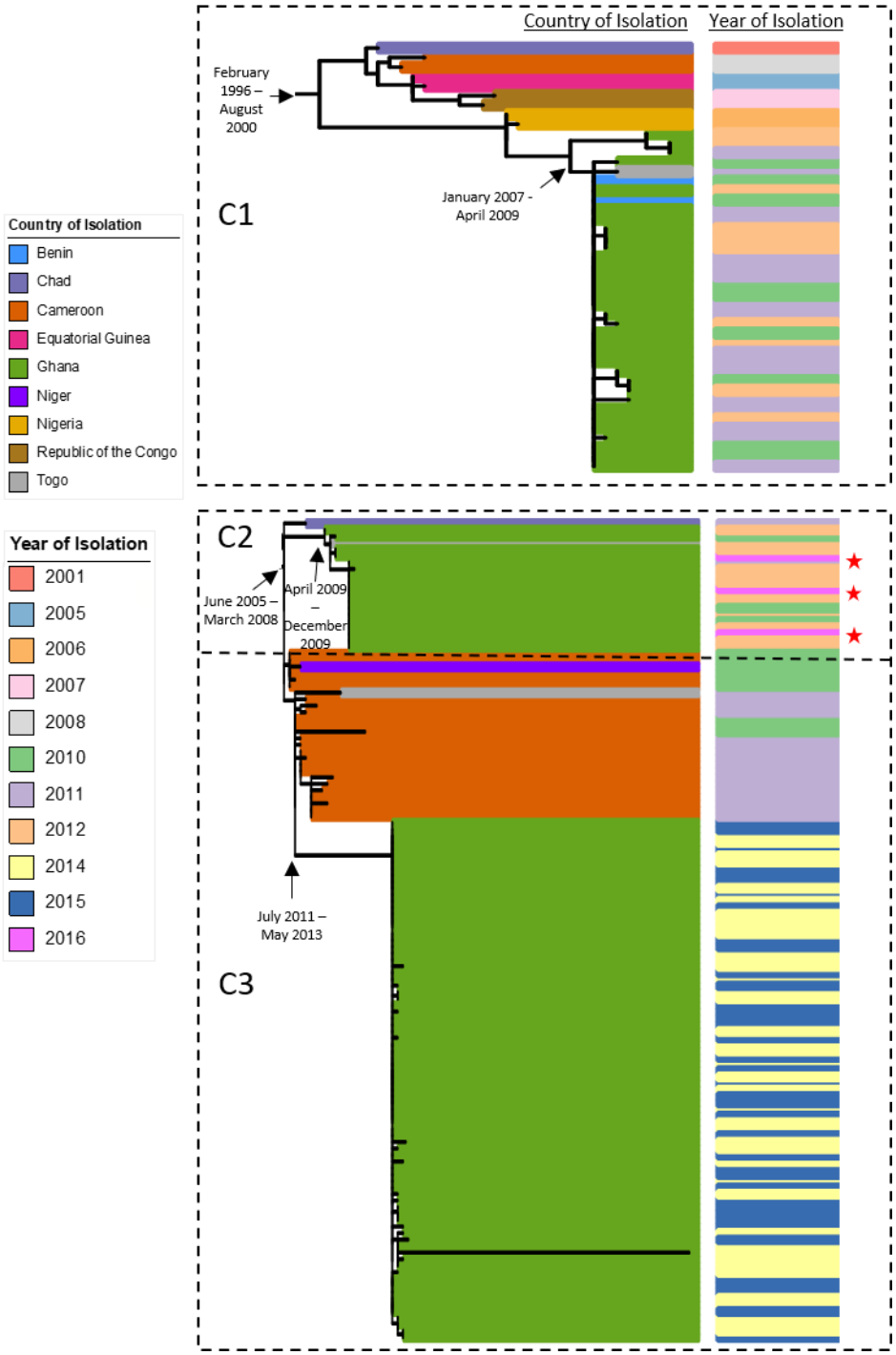

***Figure 5.52**: Subtrees of Ghanaian V. cholerae Clades 1, 2 and 3, trimmed from the maximum likelihood phylogeny in Figure 5.2, as shown in Figure 5.4. The estimated introduction timings using the 95% confidence interval are plotted, showing the ranges of the last common ancestor of Clade 1 as well as Clades 2 and 3, alongside the estimated instruction timings of each Clade to Ghana.*

5.5 Antimicrobial resistance in Ghanaian *V. cholerae* Clades

Phenotypic testing for antibiotic resistances was conducted by collaborators in Ghana, and these results showed stable antibiotic resistance profiles that were diversifying over time. All isolates tested were multidrug resistant, showing resistance to at least two antibiotic classes each. Resistance to erythromycin and trimethoprim/sulfamethoxazole, often used as frontline antibiotic treatments, were observed in *V. cholerae* isolated across all years of our collection. Resistances to cefotaxime, fluoroquinolones, and tetracyclines were rare in these isolates, before becoming much more common in those isolates from Clade 3, isolated in 2014 and 2015. Our environmental isolates showed a much higher level of diversity in the classes they were resistant to compared to most other isolates within our Ghanaian collection. All three were tested for and found to be resistant to ciprofloxacin, erythromycin, nalidixic acid, trimethoprim/sulfamethoxazole, and tetracycline, with two isolates also being resistant to chloramphenicol, and one additionally being resistant to cefotaxime.

Figure 5.6 shows the results of the *in-silico* AMR presence/absence testing for Clades 1, 2 and 3. Due to contamination identified within the nucleotide sequencing files of some of our Ghanaian *V. cholerae* isolates, the task of genotypic AMR gene presence/absence analysis was not as straight forward as it had been for *C. diphtheriae*. While any contamination has minimal impact on mapping-based approaches, the presence of contaminant DNA can present incorrect results when identifying AMR genes present within the genomic sequence files. To ensure we undertook the most comprehensive AMR analysis we could, we utilised a combination of widely used tools; Kraken2, KrakenTools and Abacas [328–330]. These methods allowed us to use multiple approaches to mask and remove reads that were not *V. cholerae*,

providing the most reliable basis to undertake phenotypic analysis of the presence and absence of AMR determinants. To do so, we used both SRST2 and ARIBA in parallel to ensure the most reliable results.

Six AMR determinant genes were found across all three distinct Clades – *strA*, *strB*, *catB5*, *floR*, *sul2* and *dfrA1*. These genes were found represented across all years of isolation, as well as all countries within the three Clades. The antibiotics that these genes are known to confer resistance to include streptomycin (*strA*, *strB*), florfenicol/chloramphenicol (*floR*, *catB5*), sulfonamide (*sul2*), and trimethoprim/sulfamethoxazole (*dfrA1* and *sul2*).

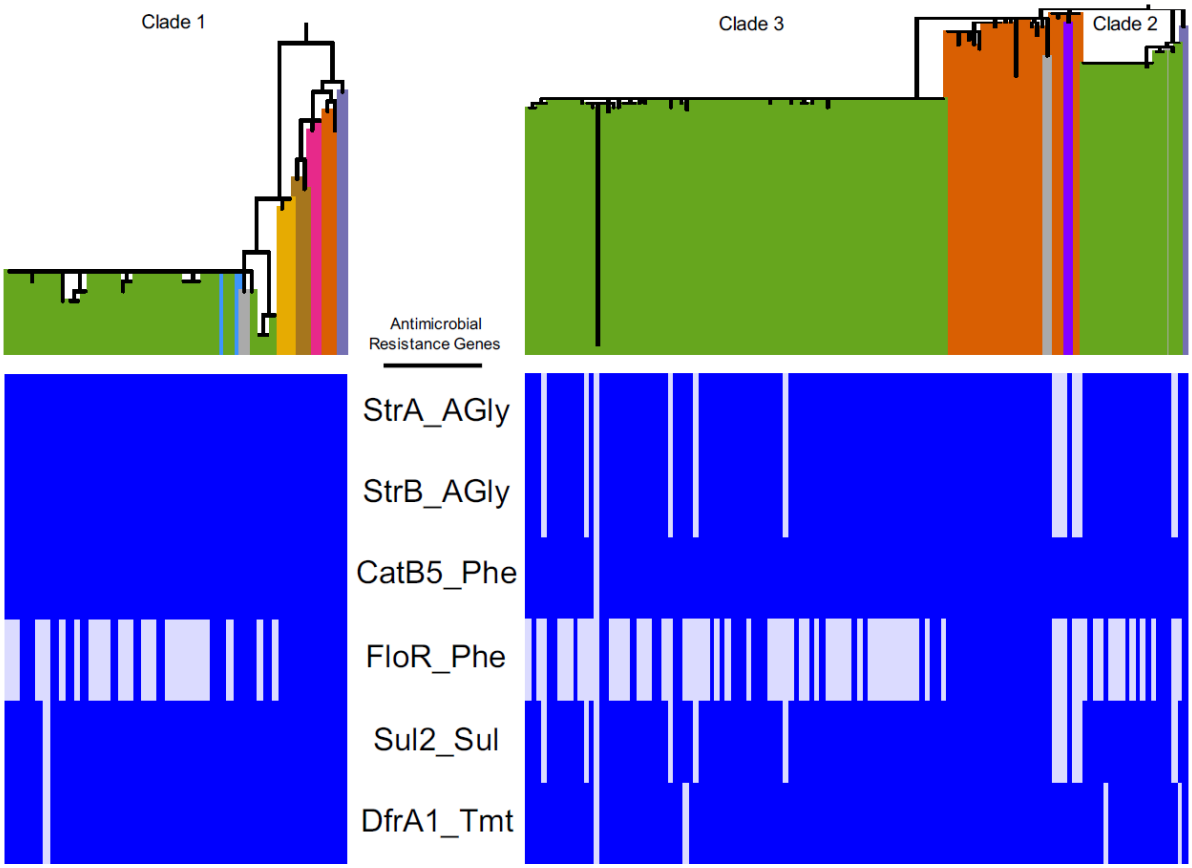

***Figure 5.53**: The presence and absence of AMR determinant genes identified by genotypic analysis across Ghanaian V. cholerae Clades 1, 2 and 3. The classes of antibiotic that each gene confers resistance to are shown next to the gene's name: AGly = aminoglycosides, Phe = phenicols, Sul = sulfonamides, Tmt = trimethoprim.*

5.6 Toxin variation and quinolone resistance mutations in Ghanaian *V. cholerae*

In-silico PCR was used to identify the variants of the cholera toxin gene *ctxB* carried by our novel Ghanaian *V. cholerae* isolates, using the same methodology as the one which we used to identify *tox* gene variants if *C. diphtheriae*. The results were consistent with which of the Clades each isolate was from, with all Clade 1 genomes exclusively harbouring the *ctxB_1* cholera toxin gene variant. Clades 2 and 3 meanwhile all harboured the *ctxB_7* variant within their genomes.

Quinolone resistance has previously been identified as a key part of the 7th Pandemic of *V. cholerae*, and we used the online CholeraeFinder tool to identify which, if any, mutations were carried by our novel Ghanaian genomes. The S83I mutation in *gyrA* and the S85L mutation in the *parC* gene were found across all isolates within Clades 2 and 3, while no mutations were identified among ant of the isolates within Clade 1.

5.6 Discussion

By bringing together the novel genome sequences of 127 Ghanaian *V. cholerae* isolates and 509 global representatives, including the four previously published isolates from Ghana in 1970 and 1971, we were able to understand how our novel genomes, and by extension Ghanaian cholera cases, fit into the global picture of the 7th *V. cholerae* Pandemic. While the four previously published Ghanaian isolates mapped within Wave 1, our 2010 – 2016 Ghanaian isolates mapped to Wave 3, alongside representatives from local and neighbouring countries in Africa. They separated into three Clades, separated temporally. While Clade 1 and 2 *V. cholerae* were isolated between 2010 and 2012, Clade 3 spanned 2014 – 2015. The recent large outbreak of cholera in Ghana, which began in 2014 and caused over 50,000 cases, is represented in Clade 3, and the sources of that outbreak were estimated to have been introduced into the country between July 2011 and May 2013. This lag period between the estimated time of introduction into Ghana and cases being reported was present in both Clades 1 and 2 as well, where introductions were estimated have happened between January 2007 – April 2009 and April – December 2009 respectively. This reinforces the belief that lag periods in cholera endemic countries is unlikely to be a real absence of the pathogen, but

rather a temporary absence of reported cholera cases due to a lack of major outbreaks in the period.

The closest relatives of our novel Ghanaian isolates were from Cameroon, Togo, Benin, Niger, and Nigeria. Recent studies, such as by Moore *et al* in 2018, have shown similar dynamics in West Africa [360]. Transmission through road and water bound trade routes have often been presented as a major transmission pathway of cholera, acting as importation routes, and Ghana is heavily linked with its surrounding local nations in such ways [360,366]. *V. cholerae* has previously been reported as first infecting coastal cities, before moving inland to more landlocked cities, often through these trade routes and the migration of people [124,178,360,361].

As we evaluate cholera in Ghana, it is important we include these aquatic environments that could be acting as reservoirs for *V. cholerae* between epidemics. Our 2016 environmental isolates clustered within Clade 2, suggesting that despite there being no subsequent cases from Clade 2 after 2012 in our collection, the Clade appears to have persisted and could indeed have remained in circulation, transmitted through contaminated water and potentially infecting asymptomatic carriers or causing cases that were not reported.

These findings can all aid in supporting targeted control of future cholera outbreak across the West African sub-region. To build from these results and obtain an even better understanding of these transmission dynamics and *V. cholerae* population structures at both a national and regional level, continuous monitoring and checking of these potential routes of importation,

both for infected individuals and potentially asymptomatic carries, to cut down on new strains being introduced into the country and causing new outbreaks. Additionally, the importance of surveying water bodies and aquatic environments for *V. cholerae* contamination is incredibly important, as previous outbreak strains remaining prevalent in the environment could quickly return under the right circumstances, such as heavy rains and floods.

Multi-drug resistance was present in all of our novel Ghanaian *V. cholerae* isolates, both in *in silico* and *in vivo* analyses, with more varied resistance profiles present among more resent isolates. It has previously been reported by Weill *et al* in 2019 that 7th Pandemic lineages resistant to multiple classes of antibiotic were introduced into Africa from Asia, where they outcompeted and replaced the existing *V. cholerae* populations [124]. Previous studies from Ghana have echoed these findings, highlighting *V. cholerae* isolates resistant to multiple classes of antibiotic, including fluoroquinolones and tetracyclines [361,367–371]. Among our collection, just under 15% of our isolates were resistant to tetracycline using WHONET breakpoints [372]. Our *in-silico* analysis showed mutations in the *gyrA* and *parC* genes among isolates from Clades 2 and 3, which correspond with another important factor in the evolution of some 7th Pandemic lineages – quinolone resistance development. We also identified resistance determinant genes encoding for streptomycin, chloramphenicol, sulfonamides, and trimethoprim/sulfamethoxazole, with five of the six genes identified by the computational tools previously reported to be carried on ICE elements, including the ubiquitous ICE*Vch*Ind5 [373,374]. Due to the steps taken to remove the contamination identified within some genomes however, our *in-silico* AMR analysis may have missed some rarer genes, and this may explain the discrepancies between our *in-silico* and *in vivo* resistance results. Future analyses of *V. cholerae* in the region may provide further evidence of tetracycline resistance development in Ghana and other neighbouring countries.

Taken collectively, our results show that *V. cholerae* isolates belonging to an individual clade are circulating in neighbouring West African countries simultaneously, and that both within these countries and across the region, multiple clades can be circulating an any given time. The extensive cross-border links, such as for trade, coupled with the large bodies of water that flow through vast swathes of populated land, provide ample opportunity for the spread and transmission of *V. cholerae*. It is important to highlight that the apparent gap in cholera cases in between outbreaks present within these countries is often not due to a lack of *V. cholerae* in the environment, but rather merely a low point in between these serious outbreaks. As the development of AMR within these strains becomes more prevalent, it is more important than ever to implement further and more continuous surveillance and control networks, to prevent the widespread movement of the pathogen between communities, and to identify the natural resource areas *V. cholerae* remains present in before future outbreaks occur.

6.0 *Vibrio cholerae* O139 serogroup

6.1 Introduction

While the *V. cholerae* O1 serogroup has been the major driver of all seven global pandemics of cholera throughout history, in 1992 a new group of Vibrio isolates causing highly similar symptoms was reported in southern and eastern India, as well as southern Bangladesh [155,375]. The variants were identified as *V. cholerae* but did not agglutinate when tested with O1 antisera [375,376]. After further testing, it was determined that the variants did not match any of the previously defined 138 O serogroups and these related isolates were named O139. Cases rapidly spiralled into large scale cholera outbreak within these countries, before spreading outwards across the region [154]. No previous serogroup except O1 had ever been shown to cause wide scale outbreaks of cholera disease, and the potential for epidemics caused by the new O139 serogroup was widely raised [375]. In many parts of Asia, including India and Bangladesh, introduced O139 serogroup isolates began to outcompete endemic O1 *V. cholerae*, following on from the initial epidemics of 1992 and 1993 through to the early 2000s [154,377]. There was even a strong belief among some researchers that O139 constituted the start of an 8th global pandemic of *V. cholerae,* that would overtake and surpass the ongoing 7th pandemic caused by *V. cholerae* O1 El Tor isolates [378].

During the emergence of O139's in 1992 and 1993, over 150,000 cholera cases caused by this serotype were reported across Bangladesh and India [379]. Analysis of these O139 serogroup isolates showed a high level of similarity to the O1 El Tor variant, in contrast to the dissimilarities between most other non-O1 serogroups including classical and El Tor O1s

[379,380]. Further phylogenetic analysis placed the new O139 isolates within Wave 2 of the 7th cholera pandemic showing they were closely related to existing O1 genomes, reinforcing the belief that O139 may have evolved from the O1 El Tor variant [120]. Evidence emerged for a significant genetic interaction between O1 and O139 Vibrios, lead to significant genomic reassortments, including in the CTX prophage, as well as in the antibiotic resistance gene repertoire in the O139 *V. cholerae* isolates [154]. Despite this rapid and dramatic emergence, O139 then vanished from epidemic settings, before sporadically remerging in outbreaks across Asia [381–383]. While O139 remained present in endemic regions such as India into the late 2000s, O1 El Tor *V. cholerae* had seemingly outcompeted the serogroup to reclaim the position of the dominant cause of cholera, both in the region and around the world [383].

A consistent question raised in the study of O139 *V. cholerae* is 'why did the serogroup emerge and then rapidly disappear? O139 *Vibrio* variants were tipped by some to be the beginning of a new global pandemic, yet within years they had almost entirely vanished from the epidemiological map, with only sporadic returns across the subsequent two decades. By bringing together a large collection of O139 isolates, we aimed to provide at least partial answers to the question. By combining representatives from across the O1 7th Pandemic, we aimed to understand the wider picture of how O139 diverged from the second wave, as well as investigating the advantages that allowed it to temporarily outcompete such a successful serogroup. We then focussed on trying to identify the genomic changes across the O139 serogroup that took the serogroup from being a major future threat to near obscurity in only a few years. I performed a preliminary analysis at the start of an international multi-disciplinary project, placing our O139 collection within the wider 7th Pandemic context, and identifying both the rapid development and subsequent equally rapid loss of antibiotic resistances across the whole population. I also defined the development of AMR within the O1 population,

presenting one potential driver of how O1 reclaimed its position as the major cause of cholera disease.

6.2 Creating a genomic collection of O139 serogroup *V. cholerae* with representative O1 isolates.

Three hundred and thirty-six novel O139 serogroup *V. cholerae* isolates were sequenced at the Wellcome Sanger Institute using an Illumina HiSeq v4 platform producing short read data. These data were combined with 31 additional O139 *V. cholerae* genomes and 258 O1 genomes from the publicly available databases National Center for Biotechnology (NCBI) Genbank and European Bioinformatics Institute (EMBL-EBI) European Nucleotide Archive to create a total collection of 625 *V. cholerae* isolates [288,318]. Our novel O139 genomes were from *Vibrios* isolated between 1992 and 2015, all from the Asian continent. The vast majority of our novel O139 were isolated in India (316), with six from Bangladesh, six from Myanmar, two from China, and the final two from Malaysia. Within the entire collection, 47 countries are represented across four continents with 65 isolates from Africa, 496 from Asia, 6 from Europe, and 55 from The Americas, while three isolates had no geographic metadata available. The genomes covered 78 years, being from isolates between 1937 and 2015. One genome represented an isolate from the 1930s and one in the 1950s, while 15 were isolated in the 1970s, 24 in the 1980s, 389 in the 1990s, 163 in the 2000s, and 17 in the 2010s. With O139 primarily reported from India and Bangladesh during the 1990s, when creating our collection of publicly available genomes we believed it was important to focus on representing both the region of Asia and the 1990s decade. We wanted to provide a local picture, while also making sure to include international representative *V. cholerae* from other regions and other time periods to provide a global and historical backdrop.

By combining this collection of novel O139 *V. cholerae* genomes with previously sequenced O139s and O1s, we hoped to further our understanding of how the O139 serogroup so successfully outcompeted the O1 serogroup during its initial emergence as a proposed 8th pandemic beginning, before providing evidence as to a potential cause of the seemingly extremely rapid decline.

6.3 The rise and fall of *V. cholerae* O139.

After assembling our 625-genome collection, construction of a *V. cholerae* phylogeny could begin. As described in the previous chapter on Ghanaian cholera, our genomes were mapped to the reference isolate N16961 using SMALT mapping, utilising the 'multiple_mapping_to_bam' and 'join_dna_with_indels' scripts produced by Simon Harris to produce a multi-genome alignment 625 genomes strong [323,324]. Gubbins was again utilised to mask recombination and to produce a SNP alignment output, in this case one 5,053 bases long. This was used as a template to produce our phylogeny.

IQ-TREE was once again utilised to produce a phylogeny over 1,000 pseudo-bootstraps, while *in silico* gene presence and absence testing was carried out to investigate the antimicrobial resistance determinant genes present within the genomes. Figure 6.1 shows the workflow depicted graphically.

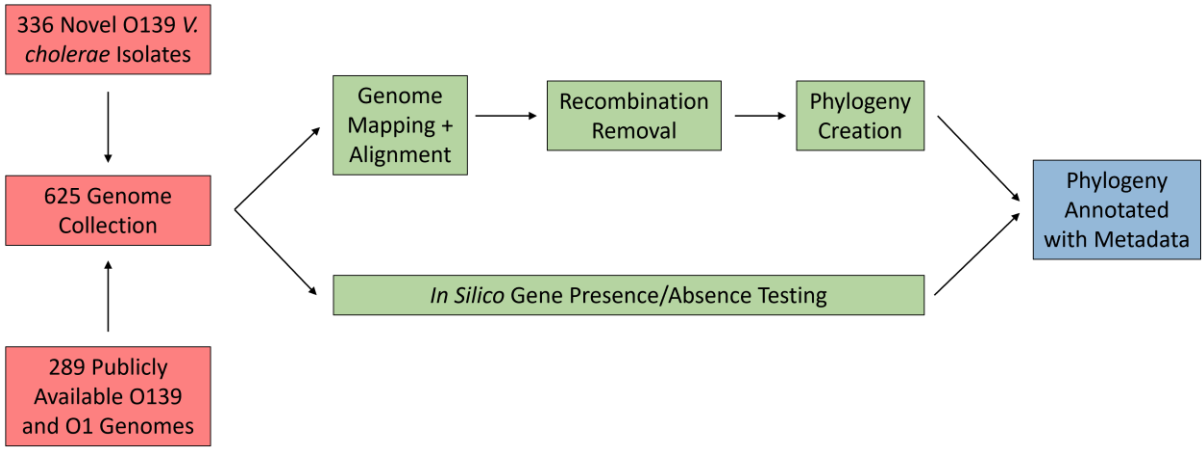

***Figure 6.54**: Flow chart diagram for generating the O139 and O1 V. cholerae isolate collection phylogeny, as well as investigating the presence and absence of AMR determinant genes. Red boxes represent the data collection phase, green the data analysis phase, and blue the data interpretation phase.*

The SNP phylogenetic tree of 625 O139 and O1 *V. cholerae* genomes (after recombination removal) can be seen in Figure 6.2. The phylogeny is annotated with the region of isolation as coloured ranges and the decade of isolation of each isolate, as well as whether the genome is O139 or O1. All O139 genomes cluster within their own separate clade, excluding one isolate that clusters within the O1 clade. All O139 *V. cholerae* – excluding the outlier – are descended from a single small cluster of O1 genomes, all isolated in 1989 and 1990, and these O1 *V. cholerae* isolates as well as their closest relatives are all part of Wave 2 of the 7th Pandemic. The O139 clade genomes represent Vibrios isolated between 1992 and 2015, with the vast majority from Asia.

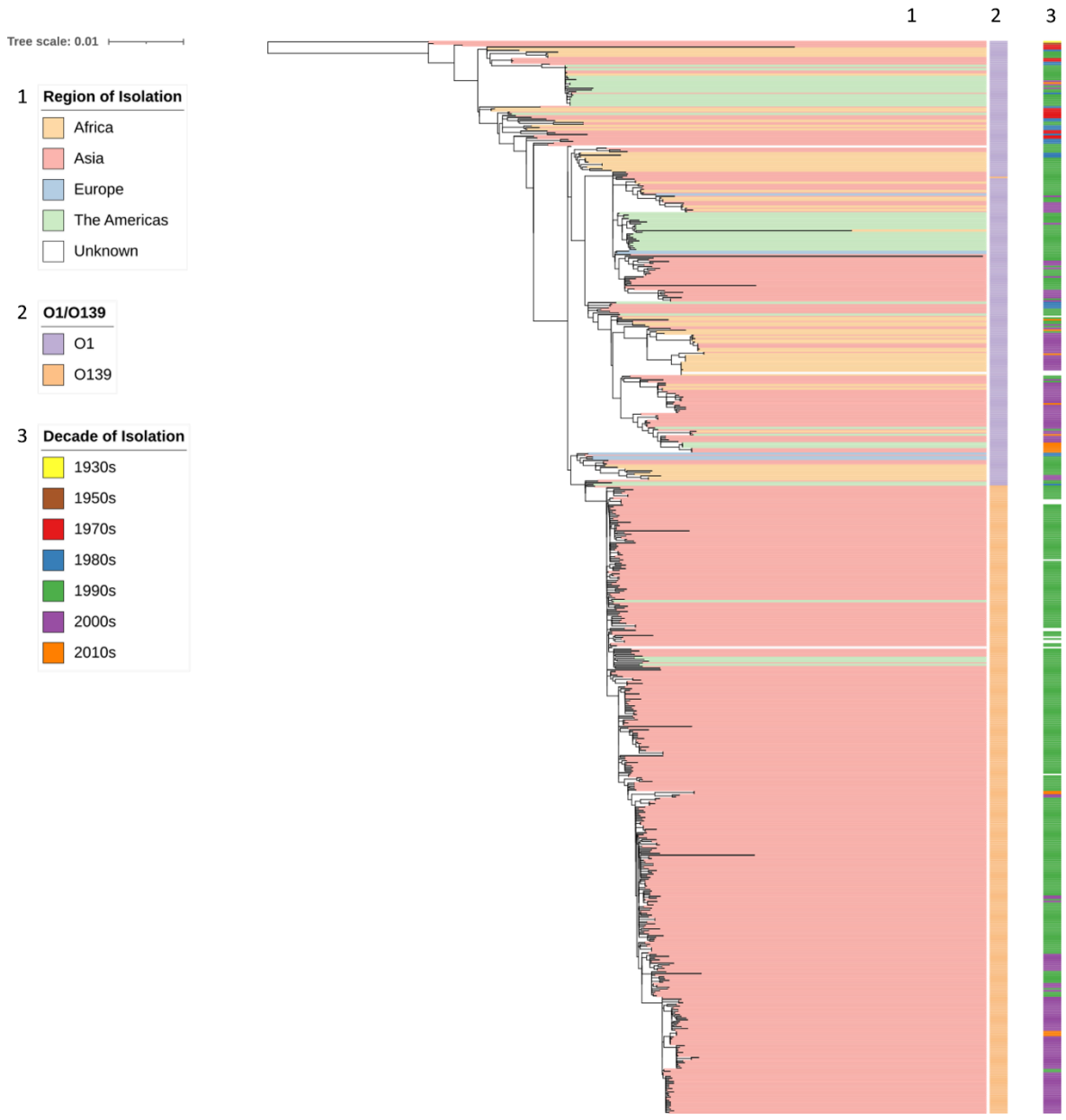

***Figure 6.55**: The maximum likelihood phylogenetic tree based on the mapped recombination-free SNPs of 625 O139 and O1 V. cholerae genome collection. The phylogeny is annotated with the region of isolation for each genome (1), whether an isolate is of the O139 or O1 serogroups (2) and the decade of isolation (3). The scale bar shows substitutions per site.*

Figure 6.3 is equivalent to Figure 6.2 but with the addition of AMR determinant genes as determined by SRST2. Multiple AMR determinant genes are present across most of the O139 *V. cholerae* genomes isolated during the 1990s, something not replicated among O1 genomes isolated during the same period. Classes of antibiotic these genes offer resistance to are

aminoglycosides (*strA* and *strB*), phenicols (*floR* and *catB5*), and sulfanomides (*sul2*). Despite this AMR profile being present in O139 *V. cholerae* isolated during the 1990s, across numerous countries in Asia, including Bangladesh, China, India, Malaysia, Myanmar, and Thailand, as well as in the Americas (Mexico), we detected a widespread loss of these genes in genomes isolated during the 2000s and 2010s. This is especially obvious among those genomes from India and Bangladesh, where the higher number of available isolates clearly demonstrate this absence, rather than it potentially being the result of a random sampling bias. Among the 324 Indian isolates, 194 of the 245 genomes (79%) isolated during the 1990s carried all five AMR genes listed above, and that number increases to 201 (82%) when those isolates with at least three genes present within their genomes are included. Among the 79 isolated in 2000 or later, only 14 (18%) carried at least three of the previously present five gene collection. In the same time period of the 2000s and 2010s that O139 isolates seemingly lost their AMR genes, O1 isolates in our collection showed the presence of these same AMR determinants genes within their genomes. In addition to the aminoglycosides (*strA* and *strB*), phenicols (*floR* and *catB5*), and sulfanomides (*sul2*) resistance profiles conferred by these genes, *dfrA1* was also present in most O1 genomes isolated during the 2000s and 2010s, which confers resistance to trimethoprim.

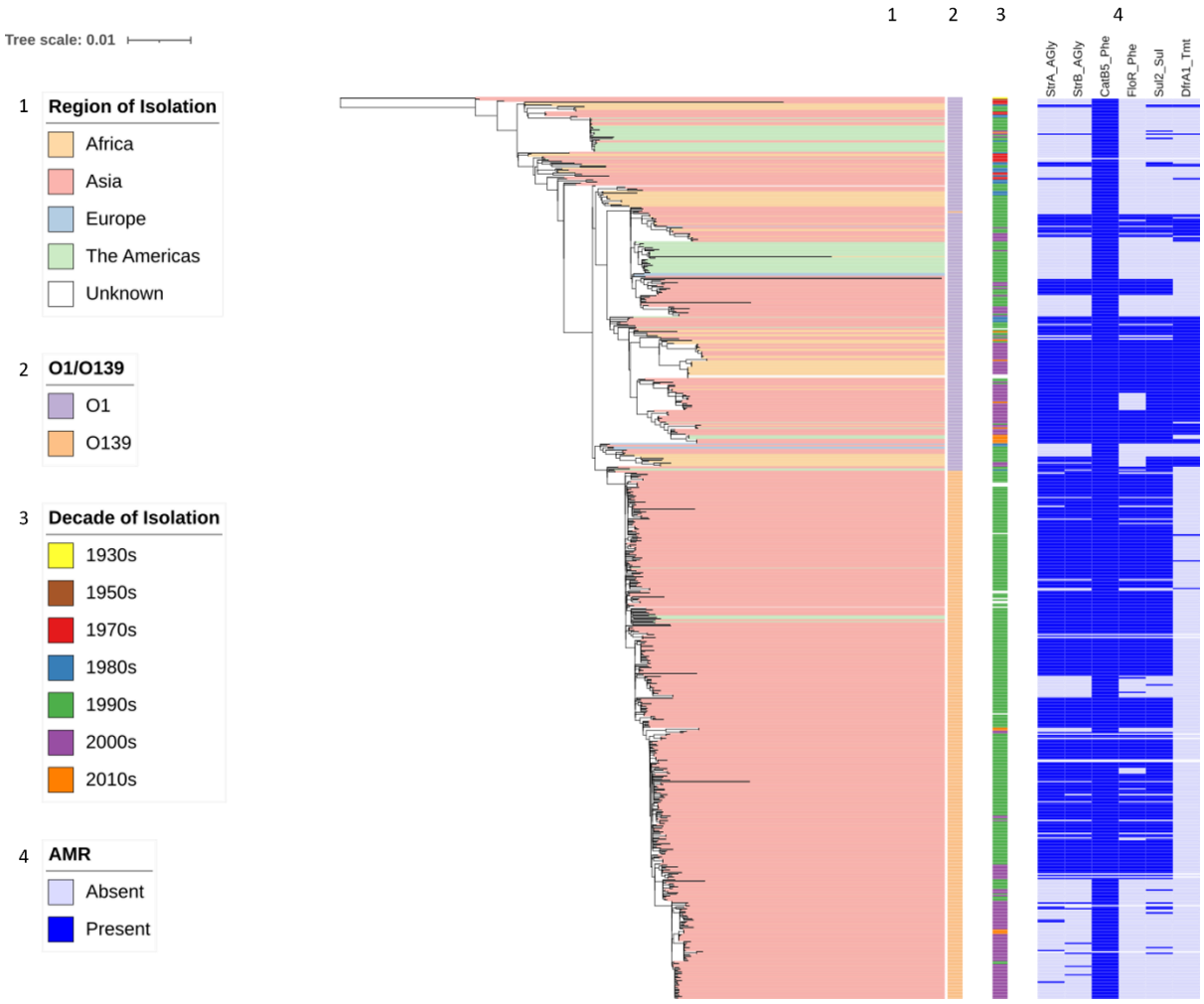

***Figure 6.56**: The maximum likelihood phylogenetic tree based on the mapped recombination-free SNPs of 625 O139 and O1 V. cholerae genome collection. The phylogeny is annotated with the region of isolation for each genome (1), whether an isolate is of the O139 or O1 serogroups (2) and the decade of isolation (3). The presence and absence of AMR determinant genes identified by genotypic analysis across the 625 genomes is also shown (4). The classes of antibiotic that each gene confers resistance to are shown next to the gene name: AGly = aminoglycosides, Phe = phenicols, Sul = sulfonamides, Tmt = trimethoprim. The scale bar shows substitutions per site.*

5.6 Discussion

By bringing together 314 newly sequenced O139 genomes and constructing a large-scale phylogeny of both O139 and O1 serogroup *V. cholerae* genomes, we can begin to present a picture of how O139 *V. cholerae* outcompeted O1 *V. cholerae* during its rapid emergence,

and how it presented the world with the potential beginning of an 8th Pandemic before disappearing almost as rapidly as it had first emerged.

Our phylogeny reaffirms the previously reported assertion that the majority of O139 serogroup isolates are all part of a single large clade that emerged from O1 El Tor isolates during the second wave of the 7th Pandemic during the late 1980s. Alongside that, the presence and absence testing of AMR genes within our isolates presented a potential reasoning as to why O139 so effectively outcompeted O1 during its emergence in the 1990s. The five AMR determinant genes present within the vast majority of O139 genomes presents a clear evolutionary advantage compared to O1 isolates present in the same locations at the same time, and could be one aspect of how O139 was able to so effectively displace existing O1 *V. cholerae* communities. The sudden loss of these AMR genes during the late 1990s and into the start of the 2000s, at the time where O1 isolates began to present not only those genes but additions as well, potentially reinforces this theory, as the O139 serogroup suffered sudden and widespread disappearances within the same time period.

The reason for this loss of AMR gene profiles can only be speculated on, and further analysis of this serogroup and any underlying genetic diversity is required to fully investigate the cause of this widespread and sudden loss, while evaluating if other factors acted to cause O139's successful emergence followed by its effective disappearance almost as quickly. Ongoing research by the O139 team will begin to present more answers to these intriguing questions.

The evolution of O139 Vibrios from an O1 El Tor sources does present a challenge for cholera surveillance, prevention, and control. Future rearrangements or mutations in the O1 El Tor lineage could result in a new serogroup that could outcompete and displaces native *V. cholerae* lineages. The emergence of any new clade might not be followed by a catastrophic widespread collapse and there is the potential to spread across the globe as a true 8th pandemic. It is thus important to continue widespread surveillance of *V. cholerae*, especially in areas where the disease is most prevalent, among both at risk populations and in environmental reservoirs. Such surveillance should provide enough time to observe any future newly adapted serogroups and lineages that are evolutionarily fitter that the ongoing O1 El Tor pandemic lineage. This would facilitate timely reactions to be made, including resource management, transmission network breaking, and vaccine development. Any such surveillance will also continue to improve our understanding of the ongoing evolution within the 7th pandemic and provide a more detailed backdrop through which to analyse *V. cholerae* globally. While the disappearance of Vibrios of the O139 serogroup is a welcome sight for the global health community, learning the lessons presented by its emergence will allow us to respond much more effectively in the future.

7.0 Future directions - lessons for outbreak preparedness

Throughout this thesis we have presented genomic analyses of two bacterial species that cause serious disease in regions across the globe. While *C. diphtheriae* and *V. cholerae* appear on the surface only distantly related, in terms of global health they share many similarities. Both are the causative agents of diseases inextricably linked to poverty, affecting those communities most at risk of serious impact by disease, socially, economically, and of course medically. Both diphtheria and cholera are easily preventable and treatable in the right facilities and with appropriate resources, but outbreaks go hand in hand with natural and man-made disasters. These results and their discussion can provide lessons for the future surveillance, response, and control of diphtheria and cholera. The increasing number of *C. diphtheriae* cases, the intriguing picture of *V. cholerae* within Ghana, and the rise and fall of O139, all present clear take-aways that are applicable not just to their own settings and fields, but to each other as well.

7.1 The persisting threat of *C. diphtheriae*

In Chapter 3, we presented a picture of the epidemiology of *C. diphtheriae* globally and national within India, the country with the highest number of cases reported yearly. Our further analysis of the *tox* gene variants present within this collection and their potential impacts on the diphtheria toxin protein were covered in Chapter 4. What has been made very clear is diphtheria as a disease is once again on the rise, with World Health Organisation numbers showing a rapid increase over the past decade [222]. Presenting a global picture of the species, built from a large collection of clinical case isolates, allowed us to begin to understand the population structure of the bacteria better than ever before.

Our collection, while assembled with a large number of publicly available genomes from across 16 countries and territories and broadly representative of the current state of global diphtheria, does have gaps, especially across Africa and the Middle East. In addition, diphtheria as a disease may be missed in settings with poor clinical infrastructure. Nevertheless, by unifying future research across the globe under a single banner with shared methods and terminologies, future analyses into the population structure of *C. diphtheriae* will not only present an even clearer picture and expand on what is contained within Chapter 3, but will also allow all subsequent research to contrast and compare their findings more easily. This is key in allowing research to rapidly be translated into impact for policy and industry.

As focus turns back towards this old disease, an ever-increasing number of genomes will become available, and this will provide a clearer scaffold through which to understand how single country and region-level outbreaks fit within the global population. This type of data is already available for other pathogen research fields, including *V. cholerae*, and provides clarity that all future analyses can be comparable. This shared infrastructure of *V. cholerae* research was utilised within Chapters 5 and 6. Our analysis of *C. diphtheriae* both within India and across the globe presented similar trends, an intriguing picture of numerous distantly related clades circulating simultaneously, and while some were present within only close geographical settings. Others were also circulating across large distances in the same years over large temporal ranges. Examples of this were highlighted by the Belarus/Germany dominated clade marked by a blue star in Figure 3.5, and in Figure 3.14's purple star-marked clade, which spanned both the North and South of India. There is a gap in the understanding

of diphtheria transmission networks at the macro-scale. There is also a lack of understanding of how *C. diphtheriae* persists in a setting without known reservoirs. Equally, it is important for future analysis of the Indian *C. diphtheriae* population structure to dive deeper into the within-state populations, establish state-to-state transmission networks, and to begin to fill in the missing states pieces of this puzzle. This more focussed analysis can also be used for sudden or ongoing outbreaks of diphtheria in India, as well as being used as a template for future analyses of other countries where diphtheria remains a serious threat. Provided resources are available to isolate and sequence genomes, outbreaks such as those that have occurred in Madagascar, The Yemen, and the Rohingya refugee settlements can now be analysed in real time, providing healthcare practitioners, researchers and policy makers with up to date information on community structure, diphtheria toxin variations, and antimicrobial resistance present among the isolates.

AMR in *C. diphtheriae* is not yet a widespread pressing concern, but the development of resistance to the numerous classes of antibiotic present within our *in-silico* testing presents several interesting questions. Coupled with the perseverance across decades of closely related *C. diphtheriae* isolates in Belarus and across Eastern Europe, the question of subsistence and reservoirs needs further analysing. Asymptomatic human carriage appears to be a strong contender theory, allowing *C. diphtheriae* to remain present within an ecosystem until it can opportunistically re-emerge. Any resident microbes may be exposed to antibiotics targeted at other bacteria, driving resistance. This could present an avenue for the development and/or uptake of AMR determinants against drugs which would not normally be used to treat diphtheria infections. Acquisition of resistance determinants by *C. diphtheriae* could be driven to the levels of these antibiotic resistance genes present within the throat microbiome. This could also be a demonstrable case of ‘collateral resistance’, as *C. diphtheriae* develops

or picks up mobile elements that offer resistance to these antibiotics 'collaterally', driven by the pressure to develop resistance against a separate antibiotic.

The key contribution of carriage remains only a theory at this time due to the lack of carriage isolates available for incorporation and testing. A widespread study, particularly among at-risk populations, to assess the presence of *C. diphtheriae* within asymptomatic individuals, is an important next step, and could provide clarity as to how the disease is so rapidly able the re-emergence under favourable conditions. Factors could include gaps in vaccination coverage and healthcare infrastructure often driven by social turmoil. While a study of this kind would have to be widespread and systematic in nature, requiring a large amount of resources and time, it could potentially fill a vital gap in our understanding of this pathogen, one that is paramount to preventing *C. diphtheriae* from continuing its return. The development of resistance to penicillin and erythromycin, while extremely rare in our collection, is incredibly important. Thus, it is essential to continue surveillance for such resistances, as both drugs remain the recommended frontline treatment options for diphtheria. Where these resistances have been reported in *C. diphtheriae*, detailed analyses such as those undertaken by Hennart *et al* who identified a novel resistance plasmid, are key to continuing our understanding of the continuing development of AMR in this pathogen [138].

The *tox* gene variants we have presented in Chapter 4 raise important questions about diphtheria control. While only one of our non-synonymous diphtheria toxin variants was estimated to have a mutation that would have a high impact on the structure of the protein, four others were estimated to have a medium impact, and one of the variants even carried two mutations, one predicted to be medium impact and the other low impact. While these

mutation impacts were estimated computationally, it is clear that further analysis into these variants *in vivo* is required to fully understand what level of impact the mutations might be having on the effectiveness of not just the diphtheria anti-toxin treatment, but also the diphtheria toxoid vaccine as well. Both present the only widely-used preventative reagents targeting the diphtheria toxin, and while there have been no reports of either becoming ineffective (nor do we present any information suggesting so), the importance of checking these facts could not be overstated. If there is an even minor effectiveness impact presented by these variants, it is critical that this be highlighted sooner rather than later, in time for actions to be taken to address them. Furthermore, further non-synonymous diphtheria toxin variants are most likely circulating now.

Future genomic analyses of *tox*$^{+}$ *C. diphtheriae* should include similar analyses to investigate the spread of not just our now-defined toxin variants, but also to identify any novel ones. Finally, while their effectiveness has not been questioned, problems with availability of the anti-toxin, as well as the requirements presented by utilising equine models for production, should highlight an opportunity for new diphtheria anti-toxin treatment development. Exploiting the quantum leaps in antibody technologies that have occurred since the development of the current anti-toxin formulation is an option. The same is true for novel vaccine development, if required in the future.

While non-toxigenic diphtheria has not been a focus of this thesis, further research is required into the mechanisms of infection by *C. diphtheriae* without the diphtheria toxin, both due to *tox*$^{-}$ genomes and those with a defunct or non-expressed *tox* gene. Indeed, this is especially required due to the rapid increase in reports of these cases over recent years. This is

especially true in areas with high vaccine coverage, and as vaccine uptake improves across the globe, the niches for non-toxigenic *C. diphtheriae* to exploit will only continue to increase.

Our analysis of *C. diphtheriae*, both across the world and within India, has presented many novel findings, from the population structure of multiple large distantly relate clades circulating simultaneously, to the AMR genes found within many recently isolated genomes. Additionally, we identified variants of the *tox* gene in *C. diphtheriae* isolated across the world. Many new questions were posed by our results however, most pressingly that of whether human carriage is playing a major role in the persistence of *C diphtheriae*, and the real-world impact that these diphtheria toxin variants may be having on the existing anti-toxin and toxoid vaccine formulations. It is paramount that these questions are investigated soon, as they are key to preventing diphtheria continuing its return towards a major global health challenge once again.

7.2 The continuing challenge of *Vibrio cholerae*

In comparison to *C. diphtheriae*, the field of *V. cholerae* has been extremely well-researched. Despite this, many gaps in our knowledge exist and questions remain, especially at national levels. In Chapter 5 we investigated the population structure of *V. cholerae* in Ghana at a local and regional level, while also presenting how the three Ghanaian clades fit within the wider 7th Pandemic. One of our key takeaways from these results was how *V. cholerae* isolates belonging to an individual clade were circulating in both Ghana and several neighbouring West African countries simultaneously, and this presents an important factor in

cholera surveillance and control going forward. Pathogens rarely respect diplomatic borders, especially in areas where cross-border migration and trade are common. Thus, it is incredibly difficult to control disease transmission within a single nation alone. In Ghana, this challenge is exacerbated by the large water bodies that flow throughout the nation, a key component in the *V. cholerae* transmission cycle. Multinational surveillance and control strategies are required to combat the disease of cholera most effectively, as any country acting alone will likely be undermined in their attempts by disease spread from areas with less effective control measures outside of their borders. International collaboration has often been a lynchpin in global health success, and the picture of cholera in West Africa is another clear example of how important that cooperation is to make real progress.

Additionally, our results highlight that the large temporal gaps in cholera cases in both Ghana and beyond should not be taken as immediate successes for public health organisations. Cholera's implicit links to poverty and natural disasters allow it to quickly re-assert itself during heavy rainy seasons and floods, something common in many regions of the world that regularly contend with cholera outbreaks. Older clades of Ghanaian *V. cholerae* appear to be subsisting in the environment for years after cases caused by those isolated are reported, and this raises the spectre of multiple clades re-emerging simultaneously if those contaminated water sources are allowed to enter human circulation. Without widespread monitoring of environmental reservoirs, successful control of human cholera cases can never be progressed to eradication, a cause that will once again require international cooperation to achieve. Worse, the progress made in cholera control being quickly reversed by these re-emergences would mean the resources spent in reaction to outbreaks would have been lost, potentially having knock-on effects that may undermine measures into the future.

The AMR profiles presented by our analysis are similar to those presented in previous studies of *V. cholerae* during the 7th pandemic, but nonetheless highlight the continued spread of determinants that make treating severe cholera that much harder [373,374]. Future analyses of Ghanaian *V. cholerae*, as well as isolates from other nations across West Africa, will present an even clearer picture of how widespread these AMR gene determinants are and how they are spreading. This adds to the concerning picture presented by *V. cholerae* isolates within the 7th Pandemic, where AMR determinants that confer resistance to many of the drugs available for treatment are becoming increasingly common. This is a trend that will seemingly continue far into the future without large-scale measures being put in place, including antibiotic stewardship.

7.3 The legacy of *Vibrio cholerae* O139

The analysis of *V. cholerae* serogroup O139 presented in Chapter 6 presents an intriguing picture with many similarities in the conclusions to those presented throughout Chapters 3 – 5. While the rapid gain and subsequent rapid loss of AMR genes presents clear evolutionary advantages and disadvantages, to say that is the sole cause of the rise and fall of O139 would be hasty. Further detailed analyses of the O139 collection are currently being undertaken by an international multi-disciplinary group of scientists, but in the meantime O139 presented an incredibly important moment for *V. cholerae* research. While the serogroup went from the potential start of an 8th Pandemic to obscurity, a future mutation of the O1 El Tor lineage could present an 8th Pandemic candidate that does not suffer a catastrophic collapse in the way serogroup O139 did. Once again, the importance of continued surveillance utilising international cooperation is paramount, as this will act as an early warning system and give

the research community time to assess the evolutionary advantages present within this new theoretical serogroup or lineage. This will also allow resources to be deployed to the right areas at the right time, and also allow the development of novel treatments when required.

7.4 Conclusion

As the world continues to grapple with the ongoing COVID-19 pandemic, gaps in our ongoing biomedical research into other diseases are naturally impacted. What *C. diphtheriae* has shown however is that even those diseases that are believed to be a relic of the past to many can quickly return as a modern problem. Equally, those diseases we consider as being well understood, such as *V. cholerae*, still have pieces of their puzzle missing, and this is evidently clear in LMICs where cases are most prevalent. As genomic technologies continue to improve and expand, the wealth of data available for analysis will increase, as will the power of the tools we use to analyse them. It is paramount we do not forget all that came before, or overlook what is happening right now outside of our local areas, lest tomorrow's challenges come through a doorway we had believed was locked for good.

References

1. Wilson, E. O. & Bossert, W. H. *Primer Of Population Biology*. (Sinauer Associates, 1971).

2. Levins, R. The Strategy Of Model Building In Population Biology. *Am. Sci.* **54**, 421–431 (1966).

3. Anderson, R. M. & May, R. M. Population biology of infectious diseases: Part I. *Nature* vol. 280 361–367 (1979).

4. May, R. M. & Anderson, R. M. Population biology of infectious diseases: Part II. *Nature* vol. 280 455–461 (1979).

5. Thieme, H. R. *Mathematics in Population Biology*. *Princeton University Press* vol. 12 (Princeton University Press, 2018).

6. Charlesworth, B. & Charlesworth, D. Population genetics from 1966 to 2016. *Hered. 2017 1181* **118**, 2–9 (2016).

7. Larson, E. J. *Evolution : the remarkable history of a scientific theory*. (Random House Publishing Group, 2006).

8. Fisher, R. A. *The genetical theory of natural selection*. (Clarendon Press, 1930).

9. Wright, S. The Distribution of Gene Frequencies Under Irreversible Mutation. *Proc. Natl. Acad. Sci. U. S. A.* **24**, 253 (1938).

10. Kimura, M. Evolutionary Rate at the Molecular Level. *Nature* **217**, 624–626 (1968).

11. Hamilton, W. D. The genetical evolution of social behaviour. II. *J. Theor. Biol.* **7**, 17–52 (1964).

12. Hamilton, W. D. Extraordinary sex ratios. *Science (80-.).* **156**, 477–488 (1967).

13. Maynard-Smith, J. Evolution and the Theory of Games. *Am. Sci.* **64**, 41–45 (1976).

14. Price, G. R. The logic of animal conflict. *Nature* **246**, 15–18 (1973).

15. Snow, J. *On the mode of communication of cholera*. (John Churchill, 1855).

16. Pietz, J., McCoy, S. & Wilck, J. H. Chasing John Snow: data analytics in the COVID-

19 era. *Eur. J. Inf. Syst.* **29**, 388–404 (2020).

17. Vineis, P. From John Snow to omics: the long journey of environmental epidemiology. *Eur. J. Epidemiol.* **33**, 355–363 (2018).

18. Didelot, X., Bowden, R., Wilson, D. J., Peto, T. E. A. & Crook, D. W. Transforming clinical microbiology with bacterial genome sequencing. *Nature Reviews Genetics* vol. 13 601–612 (2012).

19. Sanger, F., Nicklen, S. & Coulson, A. R. DNA sequencing with chain-terminating inhibitors. *Proc. Natl. Acad. Sci. U. S. A.* **74**, 5463–5467 (1977).

20. NobelPrize.org. The Nobel Prize in Chemistry 1980. *Nobel Media AB 2021* https://www.nobelprize.org/prizes/chemistry/1980/summary/ (2021).

21. Sanger, F. The early days of DNA sequences. *Nature Medicine* vol. 7 267–268 (2001).

22. Loman, N. J. & Pallen, M. J. Twenty years of bacterial genome sequencing. *Nature Reviews Microbiology* vol. 13 787–794 (2015).

23. Adams, J. DNA Sequencing Technologies. *Nat. Educ.* **1**, 193 (2008).

24. Staden, R. A strategy of DNA sequencing employing computer programs. *Nucleic Acids Res.* **6**, 2601–2610 (1979).

25. Adams, M. D. *et al.* The genome sequence of Drosophila melanogaster. *Science* vol. 287 2185–2195 (2000).

26. Sofia, H. J., Burland, V., Daniels, D. L., Plunkett, G. & Blattner, F. R. Analysis of the Escherichia coli genome. V. DNA sequence of the region from 76.0 to 81.5 minutes. *Nucleic Acids Res.* **22**, 2576–2586 (1994).

27. Levy, J. Sequencing the yeast genome: An international achievement. *Yeast* **10**, 1689–1706 (1994).

28. Glaser, P. *et al.* Bacillus subtilis genome project: cloning and sequencing of the 97 kb region from 325° to 333deg; *Mol. Microbiol.* **10**, 371–384 (1993).

29. Sulston, J. *et al.* The C. elegans genome sequencing project: A beginning. *Nature* **356**, 37–41 (1992).

30. Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science (80-.).* **269**, 496–512 (1995).

31. Fraser, C. M. *et al.* The minimal gene complement of Mycoplasma genitalium. *Science (80-.).* **270**, 397–403 (1995).

32. Goffeau, A. Life with 482 genes. *Science (80-.).* **270**, 445–446 (1995).

33. Himmelreich, R. *et al.* Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Res.* **24**, 4420–4449 (1996).

34. Blattner, F. R. *et al.* The complete genome sequence of Escherichia coli K-12. *Science* vol. 277 1453–1462 (1997).

35. Kunst, F. *et al.* The complete genome sequence of the gram-positive bacterium Bacillus subtilis. *Nature* **390**, 249–256 (1997).

36. Cole, S. T. *et al.* Deciphering the biology of mycobacterium tuberculosis from the complete genome sequence. *Nature* vol. 393 537–544 (1998).

37. Andersson, S. G. E. *et al.* The genome sequence of Rickettsia prowazekii and the origin of mitochondria. *Nature* **396**, 133–140 (1998).

38. Alm, R. A. *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen Helicobacter pylori. *Nature* **397**, 176–180 (1999).

39. Pizza, M. *et al.* Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science (80-.).* **287**, 1816–1820 (2000).

40. Moxon, R., Reche, P. A. & Rappuoli, R. Editorial: Reverse Vaccinology. *Front. Immunol.* **10**, 2776 (2019).

41. Hayashi, T. *et al.* Complete genome sequence of enterohemorrhagic Eschelichia coli O157:H7 and genomic comparison with a laboratory strain K-12. *DNA Res.* **8**, 11–22 (2001).

42. Cole, S. T. *et al.* Massive gene decay in the leprosy bacillus. *Nature* **409**, 1007–1011 (2001).

43. Kuroda, M. *et al.* Whole genome sequencing of meticillin-resistant Staphylococcus

aureus. *Lancet* **357**, 1225–1240 (2001).

44. Gauthier, J., Vincent, A. T., Charette, S. J. & Derome, N. A brief history of bioinformatics. *Brief. Bioinform.* **20**, 1981–1996 (2019).

45. Hagen, J. B. The origins of bioinformatics. *Nature Reviews Genetics* vol. 1 231–236 (2000).

46. Watson, J. D. & Crick, F. H. C. Genetical implications of the structure of deoxyribonucleic acid. *Nature* **171**, 964–967 (1953).

47. Pauling, L., Corey, R. B. & Branson, H. R. The structure of proteins; two hydrogen-bonded helical configurations of the polypeptide chain. *Proc. Natl. Acad. Sci. U. S. A.* **37**, 205–211 (1951).

48. Pauling, L. & Corey, R. B. Two Pleated-Sheet Configurations of Polypeptide Chains Involving Both Cis and Trans Amide Groups. *Proc. Natl. Acad. Sci.* **39**, 247–252 (1953).

49. Sanger, F. Chemistry of insulin. *Science (80-.).* **129**, 1340–1344 (1959).

50. Gamow, G., Rich, A. & Ycas, M. The problem of information transfer from the nucleic acids to proteins. *Adv. Biol. Med. Phys.* **4**, 23–68 (1956).

51. Ouzounis, C. A. & Valencia, A. Early bioinformatics: the birth of a discipline-a personal view. *Bioinforma. Rev.* **19**, 2176–2190 (2003).

52. McNeill, L. How Margaret Dayhoff Brought Modern Computing to Biology. *Smithsonian Magazine* https://www.smithsonianmag.com/science-nature/how-margaret-dayhoff-helped-bring-computing-scientific-research-180971904/ (2019).

53. Dayhoff, M. *Atlas of protein sequence and structure*. vol. 4 (National Biomedical Research Foundation, 1969).

54. IUPAC-IUB Commission on Biochemical Nomenclature. A one-letter notation for amino acid sequences. Tentative rules. *Biochemistry* **7**, 2703–2705 (1968).

55. Levinthal, C. Molecular model-building by computer. *Sci. Am.* **214**, 42–52 (1966).

56. Fitch, W. M. & Margoliash, E. Construction of phylogenetic trees. *Science* vol. 155

279–284 (1967).

57. Dayhoff, M. O. Computer analysis of protein evolution. *Sci. Am.* **221**, 86–95 (1969).

58. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).

59. Dayhoff, M. O., Schwartz, R. M. & Orcutt, B. C. A model of evolutionary change in proteins. in *Atlas of Protein Sequence and Structure* (National Biomedical Research Foundation, 1978).

60. Feng, D. F. & Doolittle, R. F. Progressive sequence alignment as a prerequisitetto correct phylogenetic trees. *J. Mol. Evol.* **25**, 351–360 (1987).

61. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380 (2005).

62. Roche Diagnostics GmbH. Genome Sequencer 20 System. 1–40 www.roche-applied-science.com (2006).

63. Applied Biosystems. Applied Biosystems 3500/3500xL Genetic Analyzer User Guide. (2010).

64. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).

65. Wellcome Sanger Institute. Who we are – Wellcome Sanger Institute. https://www.sanger.ac.uk/about/who-we-are/ (2021).

66. Baker, S. *et al.* High-throughput genotyping of Salmonella enterica serovar Typhi allowing geographical assignment of haplotypes and pathotypes within an urban district of Jakarta, Indonesia. *J. Clin. Microbiol.* **46**, 1741–1746 (2008).

67. National Notifiable Diseases Surveillance System (NNDSS). Salmonella Typhi Infection (Salmonella Enterica Serotype Typhi) | 2019 Case Definition. *Centers for Disease Control and Prevention* https://wwwn.cdc.gov/nndss/conditions/Salmonella-Typhi-Infection/case-definition/2019/ (2019).

68. Roumagnac, P. *et al.* Evolutionary history of Salmonella Typhi. *Science (80-.).* **314**,

1301–1304 (2006).

69. Holt, K. E. *et al.* High-throughput sequencing provides insights into genome variation and evolution in Salmonella Typhi. *Nat. Genet.* **40**, 987–993 (2008).

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

71. Van Puyvelde, S. *et al.* An African Salmonella Typhimurium ST313 sublineage with extensive drug-resistance and signatures of host adaptation. *Nat. Commun.* **10**, 1–12 (2019).

72. Holt, K. E. *et al.* Genomic analysis of diversity, population structure, virulence, and antimicrobial resistance in Klebsiella pneumoniae, an urgent threat to public health. *Proc. Natl. Acad. Sci. U. S. A.* **112**, E3574–E3581 (2015).

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

74. Voelkerding, K. V, Dames, S. A. & Durtschi, J. D. Next-Generation Sequencing: From Basic Research to Diagnostics. *Clin. Chem.* **55**, 641–658 (2009).

75. Nyrén, P., Pettersson, B. & Uhlén, M. Solid phase DNA minisequencing by an enzymatic luminometric inorganic pyrophosphate detection assay. *Anal. Biochem.* **208**, 171–175 (1993).

76. Ronaghi, M., Karamohamed, S., Pettersson, B., Uhlén, M. & Nyrén, P. Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.* **242**, 84–89 (1996).

77. Ronaghi, M., Uhlén, M. & Nyrén, P. A sequencing method based on real-time pyrophosphate. *Science* vol. 281 363–365 (1998).

78. QIAGEN. Pyrosequencing Technology and Platform Overview. https://www.qiagen.com/us/service-and-support/learning-hub/technologies-and-research-topics/pyrosequencing-resource-center/technology-overview/ (2021).

79. Cambridge Enterprise. Solexa: second-gen genetic sequencing . *Case Study*

https://www.enterprise.cam.ac.uk/case-studies/solexa-second-generation-genetic-sequencing/.

80. Illumina. Next-Generation Sequencing Basics. *A beginner's guide to NGS* https://www.illumina.com/science/technology/next-generation-sequencing/beginners.html (2021).

81. Clark, D., Pazdernik, N. & McGehee, M. *Molecular Biology*. (2018).

82. PacBio. SMRT Sequencing. https://www.pacb.com/smrt-science/smrt-sequencing/ (2021).

83. Eid, J. *et al.* Real-time DNA sequencing from single polymerase molecules. *Science (80-.).* **323**, 133–138 (2009).

84. Wu, H.-C., Astier, Y., Maglia, G., Mikhailova, E. & Bayley, H. Protein Nanopores with Covalently Attached Molecular Adapters. *J. Am. Chem. Soc.* **129**, 16142–16148 (2007).

85. Astier, Y., Braha, O. & Bayley, H. Toward single molecule DNA sequencing: Direct identification of ribonucleoside and deoxyribonucleoside 5′-monophosphates by using an engineered protein nanopore equipped with a molecular adapter. *J. Am. Chem. Soc.* **128**, 1705–1710 (2006).

86. Hoenen, T. *et al.* Nanopore sequencing as a rapidly deployable Ebola outbreak tool. *Emerg. Infect. Dis.* **22**, 331–334 (2016).

87. Quick, J. *et al.* Real-time, portable genome sequencing for Ebola surveillance. *Nature* **530**, 228–232 (2016).

88. Van Puyvelde, S. & Argimon, S. Sequencing in the time of Ebola. *Nature Reviews Microbiology* vol. 17 5 (2019).

89. Quick, J. *et al.* Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* **12**, 1261–1276 (2017).

90. Mahmoud, M., Zywicki, M., Twardowski, T. & Karlowski, W. M. Efficiency of PacBio long read correction by 2nd generation Illumina sequencing. *Genomics* **111**,

43–49 (2019).

91. Antipov, D., Korobeynikov, A., McLean, J. S. & Pevzner, P. A. hybridSPAdes: an algorithm for hybrid assembly of short and long reads. *Bioinformatics* **32**, 1009–1015 (2016).

92. Utturkar, S. M. *et al.* Evaluation and validation of de novo and hybrid assembly techniques to derive high-quality genome sequences. *Bioinformatics* **30**, 2709–2716 (2014).

93. Jeanmougin, F., Thompson, J. D., Gouy, M., Higgins, D. G. & Gibson, T. J. Multiple sequence alignment with Clustal X. *Trends Biochem. Sci.* **23**, 403–405 (1998).

94. Sievers, F. & Higgins, D. G. Clustal Omega. *Curr. Protoc. Bioinforma.* **48**, 3.13.1-3.13.16 (2014).

95. Darling, A. C. E., Mau, B., Blattner, F. R. & Perna, N. T. Mauve: Multiple alignment of conserved genomic sequence with rearrangements. *Genome Res.* **14**, 1394–1403 (2004).

96. Higgins, D. G. & Sharp, P. M. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. *Gene* **73**, 237–244 (1988).

97. Thompson, J. D., Gibson, T. J., Plewniak, F., Jeanmougin, F. & Higgins, D. G. The CLUSTAL X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* **25**, 4876–4882 (1997).

98. Higgins, D. G., Bleasby, A. J. & Fuchs, R. CLUSTAL V: Improved software for multiple sequence alignment. *Bioinformatics* **8**, 189–191 (1992).

99. Higgins, D., Sievers, F., Dineen, D. & Wilm, A. Clustal W and Clustal X Multiple Sequence Alignment. http://www.clustal.org/clustal2/ (2014).

100. Van Noorden, R., Maher, B. & Nuzzo, R. The top 100 papers. *Nature* (2014) doi:10.1038/514550a.

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

102. Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and

Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).

103. Huelsenbeck, J. P. & Crandall, K. A. Phylogeny estimation and hypothesis testing using maximum likelihood. *Annual Review of Ecology and Systematics* vol. 28 437–466 (1997).

104. Hunt, M. *et al.* ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genomics* **3**, (2017).

105. Inouye, M. *et al.* SRST2: Rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med.* **6**, 90 (2014).

106. Harris, S. R. GitHub - simonrharris/in_silico_pcr: Script to run an in silico pcr for a set of primer pairs on an assembly. *GitHub* https://github.com/simonrharris/in_silico_pcr (2016).

107. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8**, 28–36 (2017).

108. Yu, G. ggtree: Elegant Graphics for Phylogenetic Tree Visualization and Annotation. https://guangchuangyu.github.io/ggtree-book/short-introduction-to-r.html (2021).

109. RStudio Team. RStudio: Integrated development for R. *RStudio, Inc., Boston, MA* rstudio.com (2015).

110. R Core Team. R: A language and environment for statistical computing. *Vienna, Austria R Found. Stat. Comput.* 3–36 (2014).

111. Wickham, H. ggplot2. *Wiley Interdiscip. Rev. Comput. Stat.* **3**, 180–185 (2011).

112. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245 (2016).

113. Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature* **406**, 477–483 (2000).

114. Yamaichi, Y., Iida, T., Park, K. S., Yamamoto, K. & Honda, T. Physical and genetic

map of the genome of Vibrio parahaemolyticus: Presence of two chromosomes in Vibrio species. *Mol. Microbiol.* **31**, 1513–1521 (1999).

115. Trucksis, M., Michalski, J., Deng, Y. K. & Kaper, J. B. The Vibrio cholerae genome contains two unique circular chromosomes. *Proc. Natl. Acad. Sci. U. S. A.* **95**, 14464–14469 (1998).

116. Kirkup, B. C., Chang, L., Chang, S., Gevers, D. & Polz, M. F. Vibrio chromosomes share common history. *BMC Microbiol.* **10**, 1–13 (2010).

117. Chun, J. *et al.* Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 15442–15447 (2009).

118. Chin, C.-S. *et al.* The Origin of the Haitian Cholera Outbreak Strain. *N. Engl. J. Med.* **364**, 33–42 (2011).

119. Vezzulli, L., Pruzzo, C., Huq, A. & Colwell, R. R. Environmental reservoirs of Vibrio cholerae and their role in cholera. *Environ. Microbiol. Rep.* **2**, 27–33 (2010).

120. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–5 (2011).

121. Shah, M. A. *et al.* Genomic epidemiology of vibrio cholerae O1 associated with floods, Pakistan, 2010. *Emerg. Infect. Dis.* **20**, 13–20 (2014).

122. Ahmad, N. & Lodrick, D. O. Indus River. *Encyclopaedia Britannica* https://www.britannica.com/place/Indus-River (2019).

123. Domman, D. *et al.* Integrated view of Vibrio cholerae in the Americas. *Science (80-.).* **358**, 789–793 (2017).

124. Weill, F. X. *et al.* Genomic history of the seventh pandemic of cholera in Africa. *Science (80-.).* **358**, 785–789 (2017).

125. Weill, F. X. *et al.* Genomic insights into the 2016–2017 cholera epidemic in Yemen. *Nature* **565**, 230–233 (2019).

126. Cerdeño-Tárraga, A. M. *et al.* The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129. *Nucleic Acids Res.* **31**, 6516–6523 (2003).

127. Sangal, V. & Hoskisson, P. A. Evolution, epidemiology and diversity of Corynebacterium diphtheriae: New perspectives on an old foe. *Infect. Genet. Evol.* **43**, 364–370 (2016).

128. Hoskisson, P. A. Microbe Profile: Corynebacterium diphtheriae-an old foe always ready to seize opportunity. *Microbiology* (2018) doi:10.1099/mic.0.000627.

129. Trost, E. *et al.* Pangenomic Study of Corynebacterium diphtheriae That Provides Insights into the Genomic Diversity of Pathogenic Isolates from Cases of Classical Diphtheria, Endocarditis, and Pneumonia. *J. Bacteriol.* **194**, 3199–3215 (2012).

130. Grosse-Kock, S. *et al.* Genomic analysis of endemic clones of toxigenic and non-toxigenic Corynebacterium diphtheriae in Belarus during and after the major epidemic in 1990s. *BMC Genomics* **18**, 873 (2017).

131. Lodeiro-Colatosti, A. *et al.* Diphtheria Outbreak in Amerindian Communities, Wonken, Venezuela, 2016-2017. *Emerg. Infect. Dis.* **24**, 1340–1344 (2018).

132. Dangel, A., Berger, A., Konrad, R., Bischoff, H. & Sing, A. Geographically diverse clusters of nontoxigenic corynebacterium diphtheriae infection, Germany, 2016–2017. *Emerg. Infect. Dis.* **24**, 1239–1245 (2018).

133. Santos, L. S. *et al.* Diphtheria outbreak in Maranhão, Brazil: microbiological, clinical and epidemiological aspects. *Epidemiol. Infect.* **143**, 791–798 (2014).

134. Mahomed, S. *et al.* An isolated outbreak of diphtheria in South Africa, 2015. *Epidemiol. Infect.* **145**, 2100–2108 (2017).

135. Benamrouche, N. *et al.* Microbiological and molecular characterization of Corynebacterium diphtheriae isolated in Algeria between 1992 and 2015. *Clin. Microbiol. Infect.* **22**, 1005.e1-1005.e7 (2016).

136. Chorlton, S. D., Ritchie, G., Lawson, T., Romney, M. G. & Lowe, C. F. Whole-genome sequencing of corynebacterium diphtheriae isolates recovered from an inner-city population demonstrates the predominance of a single molecular strain. *J. Clin. Microbiol.* **58**, (2020).

137. Timms, V. J., Nguyen, T., Crighton, T., Yuen, M. & Sintchenko, V. Genome-wide comparison of Corynebacterium diphtheriae isolates from Australia identifies

differences in the Pan-genomes between respiratory and cutaneous strains 11 Medical and Health Sciences 1108 Medical Microbiology. *BMC Genomics* **19**, 869 (2018).

138. Hennart, M. *et al.* Population genomics and antimicrobial resistance in Corynebacterium diphtheriae. *Genome Med.* **12**, 1–18 (2020).

139. Seth-Smith, H. M. B. & Egli, A. Whole genome sequencing for surveillance of diphtheria in low incidence settings. *Frontiers in Public Health* vol. 7 235 (2019).

140. Boucher, Y., Orata, F. D. & Alam, M. The out-of-the-delta hypothesis: Dense human populations in low-lying river deltas served as agents for the evolution of a deadly pathogen. *Front. Microbiol.* **6**, 1120 (2015).

141. Shigematsu, M., Meno, Y., Misumi, H. & Amako, K. The Measurement of Swimming Velocity of Vibrio cholerae and Pseudomonas aeruginosa Using the Video Tracking Method. *Microbiol. Immunol.* **39**, 741–744 (1995).

142. Howard-Jones, N. Robert Koch and the cholera vibrio: a centenary. *Br. Med. J. (Clin. Res. Ed).* **288**, 379–381 (1984).

143. Bentivoglio, M. & Pacini, P. Filippo Pacini: A Determined Observer. *Brain Res. Bull.* **38**, 161–165 (1995).

144. Frerichs, R. R. Who first discovered cholera? *Department of Epidemiology, Fielding School of Public Health* http://www.ph.ucla.edu/epi/snow/firstdiscoveredcholera.html.

145. Das, B., Verma, J., Kumar, P., Ghosh, A. & Ramamurthy, T. Antibiotic resistance in Vibrio cholerae: Understanding the ecology of resistance genes and mechanisms. *Vaccine* vol. 38 A83–A92 (2020).

146. Almagro-Moreno, S., Pruss, K. & Taylor, R. K. Intestinal Colonization Dynamics of Vibrio cholerae. *PLoS Pathog.* **11**, e1004787 (2015).

147. Colwell, R. R. & Spira, W. M. The Ecology of Vibrio cholerae. in *Cholera* 107–127 (Springer US, 1992). doi:10.1007/978-1-4757-9688-9_6.

148. Collins, A. E. Vulnerability to coastal cholera ecology. *Soc. Sci. Med.* **57**, 1397–1407 (2003).

149. Safa, A., Nair, G. B. & Kong, R. Y. C. Evolution of new variants of Vibrio cholerae

O1. *Trends in Microbiology* vol. 18 46–54 (2010).

150. Son, M. S., Megli, C. J., Kovacikova, G., Qadri, F. & Taylor, R. K. Characterization of Vibrio cholerae O1 El tor biotype variant clinical isolates from Bangladesh and Haiti, including a molecular genetic analysis of virulence genes. *J. Clin. Microbiol.* **49**, 3739–3749 (2011).

151. Cvjetanovic, B. & Barua, D. The seventh pandemic of cholera. *Nature* **239**, 137–138 (1972).

152. Finkelstein, R. A. Cholera, Vibrio cholerae O1 and O139, and Other Pathogenic Vibrios. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 1996).

153. Cholera Working Group International Centre for Diarrhoeal Diseases Research Bangladesh. Large epidemic of cholera-like disease in Bangladesh caused by Vibrio cholerae 0139 synonym Bengal. *Lancet* **342**, 387–390 (1993).

154. Faruque, S. M. *et al.* Emergence and evolution of Vibrio cholerae O139. *Proc. Natl. Acad. Sci.* **100**, 1304–1309 (2003).

155. Faruque, S. M. *et al.* Reemergence of epidemic Vibrio cholerae O139, Bangladesh. *Emerg. Infect. Dis.* **9**, 1116–1122 (2003).

156. Chowdhury, F. *et al.* Vibrio cholerae Serogroup O139: Isolation from Cholera Patients and Asymptomatic Household Family Members in Bangladesh between 2013 and 2014. *PLoS Negl. Trop. Dis.* **9**, e0004183 (2015).

157. Harris, J. B., LaRocque, R. C., Qadri, F., Ryan, E. T. & Calderwood, S. B. Cholera. *Lancet* **379**, 2466–2476 (2012).

158. Von Seidlein, L. *et al.* The value of and challenges for cholera vaccines in Africa. *J. Infect. Dis.* **208**, (2013).

159. Richie, E. *et al.* Efficacy trial of single-dose live oral cholera vaccine CVD 103-HgR in North Jakarta, Indonesia, a cholera-endemic area. *Vaccine* **18**, 2399–2410 (2000).

160. Ferreras, E. *et al.* Single-Dose Cholera Vaccine in Response to an Outbreak in Zambia. *N. Engl. J. Med.* **378**, 577–579 (2018).

161. Qadri, F. *et al.* Efficacy of a Single-Dose, Inactivated Oral Cholera Vaccine in Bangladesh. *N. Engl. J. Med.* **374**, 1723–1732 (2016).

162. Trach, D. D. *et al.* Field trial of a locally produced, killed, oral cholera vaccine in Vietnam. *Lancet* **349**, 231–235 (1997).

163. Qadri, F. *et al.* Efficacy of a single-dose regimen of inactivated whole-cell oral cholera vaccine: results from 2 years of follow-up of a randomised trial. *Lancet Infect. Dis.* **18**, 666–674 (2018).

164. Parker, L. A. *et al.* Adapting to the global shortage of cholera vaccines: targeted single dose cholera vaccine in response to an outbreak in South Sudan. *The Lancet Infectious Diseases* vol. 17 e123–e127 (2017).

165. M'Bangombe, M. *et al.* Oral cholera vaccine in cholera prevention and control, Malawi. *Bull. World Health Organ.* **96**, 428–436 (2018).

166. World Health Organization. Cholera vaccine: WHO position paper, August 2017 – Recommendations. *Vaccine* vol. 36 3418–3420 (2018).

167. World Health Organisation. Cholera vaccines: WHO position paper. *Wkly. Epidemiol. Rec.* **34**, (2017).

168. Centers for Disease Control and Prevention. Cholera Vaccines. *National Center for Emerging and Zoonotic Infectious Diseases (NCEZID)* https://www.cdc.gov/cholera/vaccines.html (2020).

169. NHS. Cholera. https://www.nhs.uk/conditions/cholera/ (2018).

170. Bill & Melinda Gates Foundation. Enteric and Diarrheal Diseases. *Strategic Overview* https://www.gatesfoundation.org/what-we-do/global-health/enteric-and-diarrheal-diseases (2021).

171. Diaconu, K. *et al.* Cholera diagnosis in human stool and detection in water: Protocol for a systematic review of available technologies. *Syst. Rev.* **7**, 1–8 (2018).

172. Centres for Disease Control and Prevention. Cholera Illness and Symptoms. *National Center for Emerging and Zoonotic Infectious Diseases (NCEZID)* https://www.cdc.gov/cholera/illness.html (2020).

173. Hsueh, B. Y. & Waters, C. M. Combating Cholera. *F1000Research* vol. 8 1000 (2019).

174. Islam, M. T. *et al.* Field evaluation of a locally produced rapid diagnostic test for early detection of cholera in Bangladesh. *PLoS Negl. Trop. Dis.* **13**, e0007124 (2019).

175. Ramamurthy, T., Das, B., Chakraborty, S., Mukhopadhyay, A. K. & Sack, D. A. Diagnostic techniques for rapid detection of Vibrio cholerae O1/O139. *Vaccine* vol. 38 A73–A82 (2020).

176. Dick, M. H., Guillerm, M., Moussy, F. & Chaignat, C. L. Review of Two Decades of Cholera Diagnostics - How Far Have We Really Come? *PLoS Neglected Tropical Diseases* vol. 6 (2012).

177. Keddy, K. H. *et al.* Diagnosis of Vibrio cholerae O1 Infection in Africa. *J. Infect. Dis.* **208**, S23–S31 (2013).

178. Rebaudet, S., Sudre, B., Faucher, B. & Piarroux, R. Cholera in Coastal Africa: A systematic review of its heterogeneous environmental determinants. *J. Infect. Dis.* **208**, S98–S106 (2013).

179. Deen, J., Mengel, M. A. & Clemens, J. D. Epidemiology of cholera. *Vaccine* vol. 38 A31–A40 (2020).

180. World Health Organisation. Cholera, 2017. *Wkly. Epidemiol. Rec.* **93**, 489–496 (2018).

181. Mintz, E. D. & Tauxe, R. V. Cholera in Africa: A Closer Look and a Time for Action. *J. Infect. Dis.* **208**, S4–S7 (2013).

182. Qadri, F., Islam, T. & Clemens, J. D. Cholera in Yemen — An Old Foe Rearing Its Ugly Head. *N. Engl. J. Med.* **377**, 2005–2007 (2017).

183. Camacho, A. *et al.* Cholera epidemic in Yemen, 2016–18: an analysis of surveillance data. *Lancet Glob. Heal.* **6**, e680–e690 (2018).

184. Federspiel, F. & Ali, M. The cholera outbreak in Yemen: Lessons learned and way forward. *BMC Public Health* **18**, 1338 (2018).

185. Legros, D. Global Cholera Epidemiology: Opportunities to Reduce the Burden of Cholera by 2030. *J. Infect. Dis.* **218**, S137–S140 (2018).

186. Alam, M. *et al.* Diagnostic limitations to accurate diagnosis of cholera. *J. Clin. Microbiol.* **48**, 3918–3922 (2010).

187. Ali, M. *et al.* The global burden of cholera. *Bull World Heal. Organ* **90**, 209–218 (2012).

188. Devault, A. M. *et al.* Second-Pandemic Strain of Vibrio cholerae from the Philadelphia Cholera Outbreak of 1849. *N. Engl. J. Med.* **370**, 334–340 (2014).

189. Oprea, M. *et al.* The seventh pandemic of cholera in Europe revisited by microbial genomics. *Nat. Commun.* **11**, 1–10 (2020).

190. Hu, D. *et al.* Origins of the current seventh cholera pandemic. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E7730–E7739 (2016).

191. Dutilh, B. E. *et al.* Comparative genomics of 274 Vibrio cholerae genomes reveals mobile functions structuring three niche dimensions. *BMC Genomics* **15**, 654 (2014).

192. Mutreja, A. & Dougan, G. Molecular epidemiology and intercontinental spread of cholera. *Vaccine* vol. 38 A46–A51 (2020).

193. Sharma, N. C. *et al.* Diphtheria. *Nat. Rev. Dis. Prim.* **5**, 1–18 (2019).

194. Murphy, J. R. Corynebacterium Diphtheriae. in *Medical Microbiology* (ed. Baron, S.) (University of Texas Medical Branch at Galveston, 1996).

195. Loeffler, F. Untersuchungen über die Bedeutung der Mikroorganismen für die Entstehung der Diphtherie beim Menschen, bei der Traube und beim Kalbe. *Mitt KJin Gesundh* **2**, 421–499 (1884).

196. Roux, E. & Yersin, A. Contribution a l'etude de la diphtherie. *Ann. Inst. Pasteur (Paris).* **2**, 620–629 (1888).

197. Sing, A. *et al.* Corynebacterium diphtheriae in a free-roaming red fox: case report and historical review on diphtheria in animals. *Infection* **44**, 441–445 (2016).

198. Sangal, V. *et al.* A lack of genetic basis for biovar differentiation in clinically important Corynebacterium diphtheriae from whole genome sequencing. *Infect. Genet. Evol.* **21**, 54–57 (2014).

199. Dazas, M., Badell, E., Carmi-Leroy, A., Criscuolo, A. & Brisse, S. Taxonomic status of Corynebacterium diphtheriae biovar Belfanti and proposal of Corynebacterium belfantii sp. nov. *Int. J. Syst. Evol. Microbiol.* **68**, (2018).

200. Martini, H. *et al.* Diphtheria in Belgium: 2010-2017. *J. Med. Microbiol.* **68**, 1517–1525 (2019).

201. Guaraldi, A. L. de M., Hirata, R. & Azevedo, V. A. de C. Corynebacterium diphtheriae, Corynebacterium ulcerans and Corynebacterium pseudotuberculosis—General Aspects. in *Corynebacterium diphtheriae and Related Toxigenic Species* 15–37 (Springer Netherlands, 2014). doi:10.1007/978-94-007-7624-1_2.

202. Zakikhany, K. & Efstratiou, A. Diphtheria in Europe: current problems and new challenges. *Future Microbiol.* **7**, 595–607 (2012).

203. Enefer, A. *et al.* Toxigenic Corynebacterium ulcerans: re-emergence of a zoonotic infection. *Access Microbiol.* **2**, 663 (2020).

204. Sangal, V. & Hoskisson, P. A. Corynephages: Infections of the Infectors. in *Corynebacterium diphtheriae and Related Toxigenic Species* 67–81 (Springer Netherlands, 2014). doi:10.1007/978-94-007-7624-1_4.

205. Sangal, V. *et al.* Adherence and invasive properties of Corynebacterium diphtheriae strains correlates with the predicted membrane-associated and secreted proteome. *BMC Genomics* **16**, 765 (2015).

206. Zakikhany, K., Neal, S. & Efstratiou, A. Emergence and molecular characterisation of non-toxigenic tox gene-bearing Corynebacterium diphtheriae biovar mitis in the United Kingdom, 2003–2012. *Eurosurveillance* **19**, 20819 (2014).

207. Billard-Pomares, T. *et al.* Diagnosis of a non-toxigenic *tox* gene-bearing strain of *Corynebacterium diphtheriae* in a young male back from Senegal to France. *Open Forum Infect. Dis.* **4**, ofw271 (2017).

208. World Health Organization. Diphtheria vaccine: WHO position paper. *Wkly. Epidemiol. Rec.* **92**, 417–436 (2017).

209. European Centre For Disease Prevention And Control. *A case of diphtheria in Spain*. http://apps.who.int/whocc/Detail.aspx?cc_ref=UNK-194&cc_code=unk (2015).

210. De Zoysa, A. *et al.* Development, validation and implementation of a quadruplex real-time PCR assay for identification of potentially toxigenic corynebacteria. *J. Med. Microbiol.* **65**, 1521–1527 (2016).

211. Liow, Y. L. *et al. Evaluation of Conventional PCR for Detection of Toxigenic Corynebacterium diphtheriae Strains in Malaysia*. *Tropical Biomedicine* vol. 35 http://msptm.org/files/Vol35No3/775-780-Norazah-A.pdf (2018).

212. Behring & Kitasato. Ueber das Zustandekommen der Diphtherie-Immunität und der Tetanus-Immunität bei Thieren. *Dtsch. Medizinische Wochenschrift* **16**, 1113–1114 (1890).

213. NobelPrize.org. The Nobel Prize in Physiology or Medicine 1901. *Nobel Media AB 2021* https://www.nobelprize.org/prizes/medicine/1901/summary/ (2021).

214. Wenzel, E. V. *et al.* Human antibodies neutralizing diphtheria toxin in vitro and in vivo. *Sci. Rep.* **10**, 1–21 (2020).

215. Jané, M. *et al.* A case of respiratory toxigenic diphtheria: contact tracing results and considerations following a 30-year disease-free interval, Catalonia, Spain, 2015. *Euro Surveill.* **23**, 17–00183 (2018).

216. Van Damme, K. *et al.* Fatal diphtheria myocarditis in a 3-year-old girl—related to late availability and administration of antitoxin? *Paediatr. Int. Child Health* **38**, 285–289 (2018).

217. Wagner, K. S. *et al.* A review of the international issues surrounding the availability and demand for diphtheria antitoxin for therapeutic use. *Vaccine* **28**, 14–20 (2009).

218. Both, L., White, J., Mandal, S. & Efstratiou, A. Access to diphtheria antitoxin for therapy and diagnostics. *EuroSurveillance* **19**, 20830 (2014).

219. Huygen, K. Development of human monoclonal antibodies to diphtheria toxin: A solution for the increasing lack of equine DAT for therapeutic use? *Virulence* **7**, 613–615 (2016).

220. Möller, J. *et al.* Proteomics of diphtheria toxoid vaccines reveals multiple proteins that are immunogenic and may contribute to protection of humans against Corynebacterium diphtheriae. *Vaccine* **37**, 3061–3070 (2019).

221. Acosta, A. M., Moro, P. L., Hariri, S. & Tiwari, T. S. P. Diphtheria. *Pink Book* www.cdc.gov/vaccines/pubs/pinkbook/dip.html (2020).

222. World Health Organisation. Diphtheria Reported Cases. *WHO vaccine-preventable diseases: monitoring system 2020 global summary* http://apps.who.int/immunization_monitoring/globalsummary/timeseries/tsincidencediphtheria.html (2020).

223. NHS. 6-in-1 vaccine overview. https://www.nhs.uk/conditions/vaccinations/6-in-1-infant-vaccine/ (2019).

224. NHS. Diphtheria. https://www.nhs.uk/conditions/diphtheria/ (2018).

225. Clarke, K. E. N. *Review of the epidemiology of diphtheria 2000-2016*. *WHO SAGE meetings* http://www.who.int/immunization/sage/meetings/2017/april/1_Final_report_Clarke_april3.pdf?ua=1, (2017) doi:10.1371/journal.pone.0044878.

226. ECDC. European Diphtheria Surveillance Network (EDSN). *European Centre for Disease Prevention and Control* https://ecdc.europa.eu/en/about-us/partnerships-and-networks/disease-and-laboratory-networks/edsn (2019).

227. CDC. Diphtheria Surveillance. *Centers for Disease Control and Prevention* https://www.cdc.gov/diphtheria/surveillance.html (2018).

228. Filonov, V. P., Zakharenko, D. F., Vitek, C. R., Romanovsky, A. A. & Zhukovski, V. G. Epidemic Diphtheria in Belarus, 1992–1997. *J. Infect. Dis.* **181**, S41–S46 (2000).

229. Labrie, S. J., Samson, J. E. & Moineau, S. Bacteriophage resistance mechanisms. *Nat. Rev. Microbiol.* **8**, 317–327 (2010).

230. Schroven, K., Aertsen, A. & Lavigne, R. Bacteriophages as drivers of bacterial virulence and their potential for biotechnological exploitation. *FEMS Microbiol. Rev.* **45**, (2021).

231. Furfaro, L. L., Payne, M. S. & Chang, B. J. Bacteriophage Therapy: Clinical Trials and Regulatory Hurdles. *Front. Cell. Infect. Microbiol.* **8**, 376 (2018).

232. Reyes-Robles, T. *et al.* Vibrio cholerae outer membrane vesicles inhibit bacteriophage

infection. in *Journal of Bacteriology* vol. 200 (American Society for Microbiology, 2018).

233. Ackermann, H. W. Bacteriophage observations and evolution. *Res. Microbiol.* **154**, 245–251 (2003).

234. Brüssow, H. & Hendrix, R. W. Phage Genomics: Small is beautiful. *Cell* **108**, 13–16 (2002).

235. Almeida, G. M., Leppänen, M., Maasilta, I. J. & Sundberg, L. R. Bacteriophage imaging: past, present and future. *Res. Microbiol.* **169**, 488–494 (2018).

236. Van Belleghem, J., Dąbrowska, K., Vaneechoutte, M., Barr, J. & Bollyky, P. Interactions between Bacteriophage, Bacteria, and the Mammalian Immune System. *Viruses* **11**, 10 (2018).

237. Ramamurthy, T. *et al.* Revisiting the Global Epidemiology of Cholera in Conjunction With the Genomics of Vibrio cholerae. *Front. Public Heal.* **7**, 203 (2019).

238. Collier, R. J. Understanding the mode of action of diphtheria toxin: A perspective on progress during the 20th century. *Toxicon* vol. 39 1793–1803 (2001).

239. Odumosu, O., Nicholas, D., Yano, H. & Langridge, W. AB Toxins: A Paradigm Switch from Deadly to Desirable. *Toxins (Basel).* **2**, 1612 (2010).

240. Matouk, C. C. & Marsden, P. A. Molecular Insights into the Thrombotic Microangiopathies. *Mol. Genet. Basis Ren. Dis.* 453-cp4 (2008) doi:10.1016/B978-1-4160-0252-9.50030-6.

241. Friebe, S., van der Goot, F. G. & Bürgi, J. The ins and outs of anthrax toxin. *Toxins* vol. 8 69 (2016).

242. Nigam, P. & Nigam, A. Botulinum toxin. *Indian J. Dermatol.* **55**, 8–14 (2010).

243. Huang, W., Foster, J. A. & Rogachefsky, A. S. Pharmacology of botulinum toxin. *J. Am. Acad. Dermatol.* **43**, 249–259 (2000).

244. Beddoe, T., Paton, A. W., Le Nours, J., Rossjohn, J. & Paton, J. C. Structure, biological functions and applications of the AB5 toxins. *Trends in Biochemical Sciences* vol. 35 411–418 (2010).

245. Olsnes, S. & Kozlov, J. V. Ricin. *Toxicon* **39**, 1723–1728 (2001).

246. Rosenblatt-Farrell, N. The landscape of antibiotic resistance. *Environ. Health Perspect.* **117**, A244 (2009).

247. PENICILLIN'S FINDER ASSAYS ITS FUTURE; Sir Alexander Fleming Says Improved Dosage Method Is Needed to Extend Use Other Scientists Praised Self-Medication Decried. *The New York Times* 21 https://www.nytimes.com/1945/06/26/archives/penicillins-finder-assays-its-future-sir-alexander-fleming-says.html (1945).

248. Davies, J. & Davies, D. Origins and Evolution of Antibiotic Resistance. *Microbiol. Mol. Biol. Rev.* **74**, 417–433 (2010).

249. Abraham, E. P. & Chain, E. An enzyme from bacteria able to destroy penicillin. *Nature* vol. 146 837 (1940).

250. CDC. *Antibiotic Resistance Threats in the United States, 2019*. http://dx.doi.org/10.15620/cdc:82532. (2019) doi:10.15620/cdc:82532.

251. World Health Organisation. Infographic. *ANTIMICROBIAL RESISTANCE Global Report on Surveillance 2014* https://www.who.int/antimicrobial-resistance/publications/infographic-antimicrobial-resistance-20140430.pdf (2014).

252. World Health Organisation. *ANTIMICROBIAL RESISTANCE Global Report on Surveillance*. (2014).

253. Unemo, M. & Nicholas, R. A. Emergence of multidrug-resistant, extensively drug-resistant and untreatable gonorrhea. *Future Microbiology* vol. 7 1401–1422 (2012).

254. World Health Organisation. Antibiotic resistance. https://www.who.int/news-room/fact-sheets/detail/antibiotic-resistance (2020).

255. European Commission. *A European One Health Action Plan against Antimicrobial Resistance (AMR)* . http://www.who.int/entity/drugresistance/documents/surveillancereport/en/index.html (2017).

256. O'Neill, J. *Tackling drug-resistant infections globally.* (2016).

257. HM Government. *Tackling antimicrobial resistance 2019–2024 The UK's five-year national action plan*. (2019).

258. The Fleming Fund. About Us. https://www.flemingfund.org/about-us/our-aims/ (2018).

259. Lederberg, J. Cell genetics and hereditary symbiosis. *Physiol. Rev.* **32**, 403–430 (1952).

260. Macrina, F. L., Kopecko, D. J., Jones, K. R., Ayers, D. J. & McCowen, S. M. A multiple plasmid-containing Escherichia coli strain: Convenient source of size reference plasmid molecules. *Plasmid* **1**, 417–420 (1978).

261. Cheah, U. E., Weigand, W. A. & Stark, B. C. Effects of recombinant plasmid size on cellular processes in Escherichia coli. *Plasmid* **18**, 127–134 (1987).

262. Hayes, W. Recombination in Bact. coil K 12: Unidirectional transfer of genetic material. *Nature* **169**, 118–119 (1952).

263. Holloway, B. & Broda, P. William Hayes 1918-1994. *AAS Biographical Memoirs* https://www.asap.unimelb.edu.au/bsparcs/aasmemoirs/hayes.htm (1998).

264. De la Cruz, F. & Davies, J. Horizontal gene transfer and the origin of species: Lessons from bacteria. *Trends Microbiol.* **8**, 128–133 (2000).

265. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nat. Rev. Microbiol.* **3**, 679–687 (2005).

266. Thomas, C. M. & Nielsen, K. M. Mechanisms of, and barriers to, horizontal gene transfer between bacteria. *Nat. Rev. Microbiol.* **3**, 711–721 (2005).

267. Von Wintersdorff, C. J. H. *et al.* Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer. *Front. Microbiol.* **7**, 173 (2016).

268. Chen, I. & Dubnau, D. DNA uptake during bacterial transformation. *Nat. Rev. Microbiol.* **2**, 241–249 (2004).

269. Moscoso, M. & Claverys, J.-P. Release of DNA into the medium by competent Streptococcus pneumoniae: kinetics, mechanism and stability of the liberated DNA. *Mol. Microbiol.* **54**, 783–794 (2004).

270. Lorenz, M. G. & Wackernagel, W. Bacterial gene transfer by natural genetic transformation in the environment. *Microbiol. Mol. Biol. Rev.* **58**, (1994).

271. Paget, E. & Simonet, P. On the track of natural transformation in soil. *FEMS Microbiol. Ecol.* **15**, 109–117 (1994).

272. McKenna, M. Antibiotic resistance: The last resort. *Nature* vol. 499 394–396 (2013).

273. Nature. The antibiotic alarm. *Nature* **495**, 141 (2013).

274. GOV.UK. Professor Dame Sally Davies. https://www.gov.uk/government/people/sally-davies (2021).

275. Dyson, Z. A., Klemm, E. J., Palmer, S. & Dougan, G. Antibiotic Resistance and Typhoid. *Clin. Infect. Dis.* **68**, S165–S170 (2019).

276. Park, S. E. *et al.* The phylogeography and incidence of multi-drug resistant typhoid fever in sub-Saharan Africa. *Nat. Commun.* **9**, 1–10 (2018).

277. Rowe, B., Ward, L. R. & Threlfall, E. J. Multidrug-Resistant Salmonella typhi: A Worldwide Epidemic. *Clin. Infect. Dis.* **24**, S106–S109 (1997).

278. Carey, M. E. *et al.* Spontaneous Emergence of Azithromycin Resistance in Independent Lineages of *Salmonella* Typhi in Northern India. *Clin. Infect. Dis.* **72**, e120–e127 (2021).

279. Walsh, C. Where will new antibiotics come from? *Nat. Rev. Microbiol.* **1**, 65–70 (2003).

280. Roose-Amsaleg, C. & Laverman, A. M. Do antibiotics have environmental side-effects? Impact of synthetic antibiotics on biogeochemical processes. *Environ. Sci. Pollut. Res.* **23**, 4000–4012 (2016).

281. Gordillo Altamirano, F. L. & Barr, J. J. Phage therapy in the postantibiotic era. *Clinical Microbiology Reviews* vol. 32 (2019).

282. Rohde, C., Wittmann, J. & Kutter, E. Bacteriophages: A therapy concept against multi-drug-resistant bacteria. *Surg. Infect. (Larchmt).* **19**, 737–744 (2018).

283. Kortright, K. E., Chan, B. K., Koff, J. L. & Turner, P. E. Phage Therapy: A Renewed

Approach to Combat Antibiotic-Resistant Bacteria. *Cell Host and Microbe* vol. 25 219–232 (2019).

284. Ganesan, D., Gupta, S. Sen & Legros, D. Cholera surveillance and estimation of burden of cholera. *Vaccine* vol. 38 A13–A17 (2020).

285. World Health Organisation. The Global Task Force on Cholera Control. http://www.who.int/cholera/task_force/en/ (2019).

286. Mérieux Foundation. Global Task Force on Cholera Control. https://www.gtfcc.org/ (2020).

287. The Global Task Force on Cholera Control. Declaration on Ending Cholera. https://www.who.int/cholera/task_force/declaration-ending-cholera.pdf (2017).

288. World Health Organisation. *Ending Cholera A Global Roadmap To 2030*. (2017).

289. Hounmanou, Y. M. G. *et al.* Genomic insights into Vibrio cholerae O1 responsible for cholera epidemics in Tanzania between 1993 and 2017. *PLoS Negl. Trop. Dis.* **13**, e0007934 (2019).

290. Dureab, F., Müller, O. & Jahn, A. Resurgence of diphtheria in Yemen due to population movement. *J. Travel Med.* **25**, (2018).

291. Dureab, F. *et al.* Diphtheria outbreak in Yemen: The impact of conflict on a fragile health system. *Confl. Health* **13**, 19 (2019).

292. PHG Foundation. What we do. https://www.phgfoundation.org/what-we-do (2021).

293. Luheshi, L. *et al. Pathogen Genomics Into Practice*. www.phgfoundation.org (2015).

294. Chewapreecha, C. *et al.* Dense genomic sampling identifies highways of pneumococcal recombination. *Nat. Genet.* **46**, 305–309 (2014).

295. Croucher, N. J. *et al.* Population genomics of post-vaccine changes in pneumococcal epidemiology. *Nat. Genet.* **45**, 656–663 (2013).

296. Croucher, N. J., Harris, S. R., Barquist, L., Parkhill, J. & Bentley, S. D. A High-Resolution View of Genome-Wide Pneumococcal Transformation. *PLoS Pathog.* **8**, e1002745 (2012).

297. Golubchik, T. *et al.* Pneumococcal genome sequencing tracks a vaccine escape variant formed through a multi-fragment recombination event. *Nat. Genet.* **44**, 352–355 (2012).

298. Smith, G. J. D. *et al.* Origins and evolutionary genomics of the 2009 swine-origin H1N1 influenza a epidemic. *Nature* **459**, 1122–1125 (2009).

299. Gire, S. K. *et al.* Genomic surveillance elucidates Ebola virus origin and transmission during the 2014 outbreak. *Science (80-.).* **345**, 1369–1372 (2014).

300. European Centre for Disease Prevention and Control. COVID-19 situation update worldwide, as of week 8, updated 4 March 2021. https://www.ecdc.europa.eu/en/geographical-distribution-2019-ncov-cases (2021).

301. The Lancet. Genomic sequencing in pandemics. *Lancet* **397**, 445 (2021).

302. COVID-19 Genomics UK Consortium. About Us. https://www.cogconsortium.uk/cog-uk/about-us/ (2021).

303. Sharon Peacock. A short history of the COVID-19 Genomics UK (COG-UK) Consortium. *COVID-19 Genomics UK Consortium* https://www.cogconsortium.uk/a-short-history-of-the-covid-19-genomics-uk-cog-uk-consortium/ (2020).

304. William A. Haseltine. Why The U.S. Needs To Step Up Covid-19 Genome Sequencing. *Forbes* (2021).

305. Furuse, Y. Genomic sequencing effort for SARS-CoV-2 by country during the pandemic. *Int. J. Infect. Dis.* **103**, 305–307 (2021).

306. COVID-19 Genomics UK Consortium. COG UK. https://www.cogconsortium.uk/ (2021).

307. COVID-19 Genomics UK Consortium. COG-UK passes 100K genomes. https://www.cogconsortium.uk/cog-uk-passes-100k-genomes/ (2020).

308. COG-UK. Microreact - UK SARS-CoV-2. https://beta.microreact.org/project/fTBJHYg2JYmXShmwgG6Soa-cog-uk-2021-03-06-uk-sars-cov-2/?dfr=lineage&cbc=lineage (2021).

309. du Plessis, L. *et al.* Establishment and lineage dynamics of the SARS-CoV-2 epidemic

in the UK. *Science (80-.).* **371**, 708–712 (2021).

310. Pfefferbaum, B. & North, C. S. Mental Health and the Covid-19 Pandemic. *N. Engl. J. Med.* **383**, 510–512 (2020).

311. Nelson, M. I. Tracking the UK SARS-CoV-2 outbreak. *Science (80-.).* **371**, 680–681 (2021).

312. Cheng, V. C. C., Lau, S. K. P., Woo, P. C. Y. & Kwok, Y. Y. Severe acute respiratory syndrome coronavirus as an agent of emerging and reemerging infection. *Clinical Microbiology Reviews* vol. 20 660–694 (2007).

313. Talavera, A. & Pérez, E. M. Is cholera disease associated with poverty? *J. Infect. Dev. Ctries.* **3**, 408–411 (2009).

314. Raad, I. I., Chaftari, A.-M., Dib, R. W., Graviss, E. A. & Hachem, R. Emerging outbreaks associated with conflict and failing healthcare systems in the Middle East. *Infect. Control Hosp. Epidemiol.* **39**, 1230–1236 (2018).

315. Finger, F. *et al.* Real-time analysis of the diphtheria outbreak in forcibly displaced Myanmar nationals in Bangladesh. *bioRxiv* (2018) doi:10.1101/388645.

316. Rahman, M. R. & Islam, K. Massive diphtheria outbreak among Rohingya refugees: lessons learnt. *J. Travel Med.* **26**, (2019).

317. Chan, E. Y. Y., Chiu, C. P. & Chan, G. K. W. Medical and health risks associated with communicable diseases of Rohingya refugees in Bangladesh 2017. *International Journal of Infectious Diseases* vol. 68 39–43 (2018).

318. Leinonen, R. *et al.* The European nucleotide archive. *Nucleic Acids Res.* **39**, D28–D31 (2011).

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

320. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).

321. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15–e15

(2015).

322. Hadfield, J. *et al.* Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* **34**, 292–293 (2018).

323. sanger-pathogens. bact-gen-scripts. *GitHub* https://github.com/sanger-pathogens/bact-gen-scripts (2020).

324. Ponstingl, H. & Ning, Z. SMALT. *Wellcome Sanger Institute* https://www.sanger.ac.uk/tool/smalt-0/ (2010).

325. Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., Von Haeseler, A. & Jermiin, L. S. ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nat. Methods* **14**, 587–589 (2017).

326. Rambaut, A. Figtree, version 1.4.3. http://tree.bio.ed.ac.uk/software/figtree (2009).

327. Gupta, S. K. *et al.* ARG-annot, a new bioinformatic tool to discover antibiotic resistance genes in bacterial genomes. *Antimicrob. Agents Chemother.* **58**, 212–220 (2014).

328. Lu, J. KrakenTools. *Github* https://github.com/jenniferlu717/KrakenTools (2021).

329. Wood, D. E., Lu, J. & Langmead, B. Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 1–13 (2019).

330. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: Algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25**, 1968–1969 (2009).

331. Berman, H., Henrick, K. & Nakamura, H. Announcing the worldwide Protein Data Bank. *Nat. Struct. Mol. Biol.* **10**, 980–980 (2003).

332. Goddard, T. D. *et al.* UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14–25 (2018).

333. Louie, G. V., Yang, W., Bowman, M. E. & Choe, S. Crystal structure of the complex of diphtheria toxin with an extracellular fragment of its receptor. *Mol. Cell* **1**, 67–78 (1997).

334. Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N. & Sternberg, M. J. E. The Phyre2 web portal for protein modeling, prediction and analysis. *Nat. Protoc.* **10**, 845–858 (2015).

335. Yates, C. M., Filippis, I., Kelley, L. A. & Sternberg, M. J. E. SuSPect: Enhanced prediction of single amino acid variant (SAV) phenotype using network features. *J. Mol. Biol.* **426**, 2692–2701 (2014).

336. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7**, 1–8 (2007).

337. Corander, J., Waldmann, P., Marttinen, P. & Sillanpää, M. J. BAPS 2: enhanced possibilities for the analysis of genetic population structure. *Bioinformatics* **20**, 2363–2369 (2004).

338. Center for Genomic Epidemiology. CholeraeFinder 1.0. https://cge.cbs.dtu.dk/services/CholeraeFinder/ (2020).

339. Wong, V. K. *et al.* Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. *Nat. Genet.* **47**, 632–639 (2015).

340. Pearce, M. E. *et al.* Comparative analysis of core genome MLST and SNP typing within a European Salmonella serovar Enteritidis outbreak. *Int. J. Food Microbiol.* **274**, 1 (2018).

341. Google Maps. India - Google Maps. https://www.google.co.uk/maps (2020).

342. Davidson, R. M. A Closer Look at the Genomic Variation of Geographically Diverse Mycobacterium abscessus Clones That Cause Human Infection and Disease. *Front. Microbiol.* **9**, 2988 (2018).

343. Love, W. J., Zawack, K. A., Booth, J. G., Gröhn, Y. T. & Lanzas, C. Markov Networks of Collateral Resistance: National Antimicrobial Resistance Monitoring System Surveillance Results from Escherichia coli Isolates, 2004-2012. *PLoS Comput. Biol.* **12**, 1005160 (2016).

344. Wagner, K. S. *et al.* Diphtheria in the postepidemic period, Europe, 2000-2009. *Emerg. Infect. Dis.* **18**, 217–25 (2012).

345. Choe, S. *et al.* The crystal structure of diphtheria toxin. *Nature* **357**, 216–222 (1992).

346. Mateyak, M. K. & Kinzy, T. G. ADP-ribosylation of translation elongation factor 2 by diphtheria toxin in yeast inhibits translation and cell separation. *J. Biol. Chem.* **288**, 24647–24655 (2013).

347. Collier, R. J. Effect of diphtheria toxin on protein synthesis: inactivation of one of the transfer factors. *Bacteriol. Rev.* **39**, 83–98 (1975).

348. White, A., Ding, X., vanderSpek, J. C., Murphy, J. R. & Ringe, D. Structure of the metal-ion-activated diphtheria toxin repressor/ tox operator complex. *Nature* **394**, 502–506 (1998).

349. Wang, Z., Schmitt, M. P. & Holmes, R. K. *Characterization of Mutations That Inactivate the Diphtheria Toxin Repressor Gene (dtxR)*. *Infection and Immunity* vol. 62 http://iai.asm.org/ (1994).

350. Wittchen, M. *et al.* Transcriptome sequencing of the human pathogen Corynebacterium diphtheriae NCTC 13129 provides detailed insights into its transcriptional landscape and into DtxR-mediated transcriptional regulation. *BMC Genomics* **19**, 82 (2018).

351. Gouy, M., Guindon, S. & Gascuel, O. SeaView Version 4: A Multiplatform Graphical User Interface for Sequence Alignment and Phylogenetic Tree Building. *Mol. Biol. Evol.* **27**, 221–224 (2010).

352. Argimón, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology and phylogeography. *Microb. Genomics* **2**, e000093 (2016).

353. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. (2013).

354. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **41**, D36–D42 (2012).

355. World Health Organization. WHO Fact Sheet: Cholera. *World Health Organization - Fact Sheets* (2019).

356. Azman, A. S., Rudolph, K. E., Cummings, D. A. T. & Lessler, J. The incubation period of cholera: A systematic review. *J. Infect.* **66**, 432–438 (2013).

357. Vezzulli, L., Colwell, R. R. & Pruzzo, C. Ocean Warming and Spread of Pathogenic Vibrios in the Aquatic Environment. *Microb. Ecol.* **65**, 817–825 (2013).

358. Colwell R R. Global climate and infections: the cholera paradigm. *Science (80-.).* **274**, 2025–2031 (1996).

359. WHO. Global Health Atlas. *World Health Organization* http://apps.who.int/globalatlas/dataQuery/.

360. Moore, S. *et al.* Dynamics of cholera epidemics from Benin to Mauritania. *PLoS Negl. Trop. Dis.* **12**, e0006379 (2018).

361. Opintan, J. A., Newman, M. J., Nsiah-Poodoh, O. A. & Okeke, I. N. Vibrio cholerae O1 from Accra, Ghana carrying a class 2 integron and the SXT element. *J. Antimicrob. Chemother.* **62**, 929–933 (2008).

362. Ghana Health Service. *Ghana Health Service Annual Report for 2014*. *Ghana Health Service* https://www.ghanahealthservice.org/ghs-item-details.php?cid=2&scid=52&iid=107 (2014).

363. Thompson, C. C. *et al.* Vibrio cholerae O1 lineages driving cholera outbreaks during seventh cholera pandemic in Ghana. *Infect. Genet. Evol.* **11**, 1951–1956 (2011).

364. Eibach, D. *et al.* Molecular Epidemiology and Antibiotic Susceptibility of Vibrio cholerae Associated with a Large Cholera Outbreak in Ghana in 2014. *PLoS Negl. Trop. Dis.* **10**, e0004751 (2016).

365. Tonkin-Hill, G., Lees, J. A., Bentley, S. D., Frost, S. D. W. & Corander, J. RhierBAPs: An R implementation of the population clustering algorithm hierBAPS. *Wellcome Open Res.* **3**, 93 (2018).

366. Okeke, I. N. Africa in the Time of Cholera: A History of Pandemics from 1817 to the Present. *Emerg. Infect. Dis.* **18**, 362–362 (2012).

367. Marin, M. A. *et al.* Cholera Outbreaks in Nigeria Are Associated with Multidrug Resistant Atypical El Tor and Non-O1/Non-O139 Vibrio cholerae. *PLoS Negl. Trop. Dis.* **7**, e2049 (2013).

368. Kiiru, J. *et al.* A Study on the Geophylogeny of Clinical and Environmental Vibrio

cholerae in Kenya. *PLoS One* **8**, 1–8 (2013).

369. Pena, E. S., Kakaiuml, C. G., Bompangueacute, D. & Toureacute, K. Cholera: Evolution of Epidemiological Situation in four French-speaking African Countries from 2004 to 2013. *West Afr. J. Med.* **33**, 245–251 (2014).

370. Kuma, G. K. *et al.* Antibiotic resistance patterns amongst clinical Vibrio cholerae O1 isolates from Accra, Ghana Outbreak Investigation View project Viruses in Dental diseases View project. *Artic. Int. J. Infect. Control* **10**, . (2014).

371. Danso, E. K. *et al.* A molecular and epidemiological study of Vibrio cholerae isolates from cholera outbreaks in southern Ghana. *PLoS One* **15**, (2020).

372. O'Brien, T. F. & Stelling, J. WHONET – Tracking microbes for patient safety. *J. Microbiol. Immunol. Infect.* **48**, S22 (2015).

373. Ceccarelli, D., Spagnoletti, M., Bacciu, D., Cappuccinelli, P. & Colombo, M. M. New V. cholerae atypical El Tor variant emerged during the 2006 epidemic outbreak in Angola. *BMC Microbiol.* **11**, 1–8 (2011).

374. Spagnoletti, M. *et al.* Acquisition and evolution of SXT-R391 integrative conjugative elements in the seventh-pandemic Vibrio cholerae lineage. *MBio* **5**, (2014).

375. Cholera Working Group International Center for Diarrhoeal Disease Research. Large epidemic of cholera-like disease in Bangladesh caused by Vibrio cholerae O139 synonym Bengal. *Lancet* **342**, 387–390 (1993).

376. Ramamurthy, T. *et al.* Emergence of novel strain of Vibrio cholerae with epidemic potential in southern and eastern India. *Lancet* **341**, 703–704 (1993).

377. Nair, G. B. *et al.* Spread of vibrio cholerae 0139 bengal in india. *J. Infect. Dis.* **169**, 1029–1034 (1994).

378. Nair, Bhattacharya, S. & Deb, B. Vibrio cholerae O139 Bengal : the eighth pandemic strain of cholera. *Indian J. Public Health* **38**, 33 (2021).

379. Popovic, T. *et al.* Molecular Subtyping Of Toxigenic Vibrio Cholerae O139 Causing Epidemic Cholera In India And Bangladesh, 1992–1993. *J. Infect. Dis.* **171**, 122–127 (1995).

380. Berche, P. *et al.* The Novel Epidemic Strain 0139 Is Closely Related To The Pandemic Strain 01 Of Vibrio Cholerae [X]. *J. Infect. Dis.* **170**, 701–704 (1994).

381. MJ, A. Epidemiology & molecular biology of Vibrio cholerae O139 Bengal. *Indian J. Med. Res.* **104**, 14–27 (1996).

382. Faruque, S. M. *et al.* Reemergence of Epidemic Vibrio cholerae O139, Bangladesh. *Emerg. Infect. Dis.* **9**, 1116 (2003).

383. Ghosh, R. *et al.* Phenotypic and Genetic Heterogeneity in Vibrio cholerae O139 Isolated from Cholera Cases in Delhi, India during 2001–2006. *Front. Microbiol.* **0**, 1250 (2016).