Chapter 1

Introduction

1.1 – Initial overview: Cholera, cholera incidence, and case definitions

1.1.1 – Clinical presentation and incidence

Cholera is an acute diarrhoeal disease transmitted *via* the faecal-oral route, caused by the Gram-negative bacterium *Vibrio cholerae* [1–3]. Cholera leads to very rapid dehydration, and is characterised by the profuse production of 'rice-water' stools flaked with mucus [2]. Incomplete international reporting of cholera cases means that estimates of the global burden of cholera disease are underestimates [4]. Nonetheless, it has been estimated that there are 2.8-2.9 million cases of cholera annually [4, 5], and up to 143,000 deaths from cholera *per annum* [4, 6]. Without intervention, global cholera cases have been projected to rise to ~3.7 million cases *per annum* by 2030 [7]. It has also been estimated that 1.3 billion people in endemic countries are at risk of contracting cholera [4]. Cholera is considered by some to be a disease of the poor, and low-income countries have been shown to be more affected by cholera than middle- or high-income countries [8]. The World Health Organisation (WHO) and Global Taskforce for Cholera Control (GTFCC) have published a strategic roadmap which aims to see cholera eliminated by the year 2030 [9].

1.1.2 – Treatment and vaccines

The principal treatment for clinical cases of cholera is the administration of rehydration therapy, either oral rehydration if no or some dehydration has occurred, or intravenous rehydration for cases of severe dehydration or shock [10]. Although cholera is caused by a bacterial pathogen, the administration of antibiotics is only recommended to occur in conjunction with, rather than instead of, oral rehydration [11]. Historically, tetracycline was the antibiotic of choice for treatment of severe cholera cases [12], and current guidance from the Centers for Disease Control and Prevention (CDC) is that doxycycline and tetracycline are useable in the treatment of cholera although azithromycin, erythromycin, and other antimicrobials are also recommended [11]. Based on data collated by the CDC [11], doxycycline is recommended as the first-line antimicrobial of choice for cholera treatment by the WHO [13], the Pan-American Health Organisation (PAHO) [14], the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b) [15], and Médecins Sans Frontières (MSF) [16].

Vaccines are available for cholera, which are principally either whole-cell/toxoid vaccines, or composed of live-attenuated vaccine strains of *V. cholerae* [1, 17], strains which may be genetically engineered [18, 19]. At the time of writing, at least seven oral killed cholera vaccines and four live-attenuated oral vaccines are either licenced or under development [20]. Three vaccines have been pre-qualified by the WHO – Dukoral®, Shanchol™ and Euvichol® – and of these, Shanchol™ and Euvichol® are included in the global oral cholera vaccine (OCV) stockpile created by WHO and funded in part by GAVI, the Global Alliance for Vaccines and Immunisation [17, 20]. This vaccine stockpile has been reactively deployed to reduce cholera transmission in settings of humanitarian crisis, such as amongst the Rohingya refugee population in Bangladesh [21]. The development of new cholera vaccines and vaccine strains is an area of active research, and strains such as HaitiV have been designed to be intrinsically recalcitrant to becoming toxigenic [22]; these strains have shown evidence of conferring protection against infection with toxigenic *V. cholerae* in mouse models [23, 24].

1.1.3 – Epidemiology and case definitions

One of the most distinct and notorious epidemiological features of cholera is the ability of its aetiological agent to spread rapidly and to cause explosive outbreaks. These transmissions and outbreaks are exemplified by the Haitian cholera epidemic of 2010, associated with an intercontinental transmission event [25–27], and more recently, outbreaks in Yemen in mid-2017 [28]. In addition to causing acute community outbreaks and national epidemics, *V. cholerae* also has the capacity to spread internationally in multi-continent cholera pandemics [1]. The epidemiology of cholera outbreaks and epidemics is traditionally associated with the work of John Snow, an English anaesthetist who mapped cholera cases in Broad Street and adjacent areas in London between 1849 and the 1850s [29, 30]. Snow's observations identified contaminated water sources as the sources of cholera transmission within and around Broad Street, and were published in 1855 [30]. These are regarded as seminal examples of epidemiological practice [31].

In order to track internationally the incidence and spread of cholera, common clinical definitions of a cholera case are required. The WHO and GTFCC, in the Roadmap to ending cholera by 2030 [9], define a confirmed cholera case as:

> "A suspected case with *Vibrio cholerae* O1 or O139 confirmed by culture or PCR polymerase chain reaction and, in countries where cholera is not present or has been eliminated, the *Vibrio cholerae* O1 or O139 strain is demonstrated to be toxigenic" [9].

The same cholera case definition is used by the CDC [3, 32]. The definition for suspected cholera cases, which is also shared by the WHO and CDC [3, 9, 32], depends on whether or not a country is defined as suffering currently from an outbreak of cholera (defined as "the occurrence of at least one confirmed case of cholera and evidence of local transmission" [9]). In countries where cholera outbreaks have not been declared, a suspected cholera case is defined as:

> "Any patient 2 years old or older presenting with acute watery diarrhea and severe dehydration or dying from acute watery diarrhea" [3, 32].

Acute watery diarrhoea (AWD) is defined as "three or more loose or watery (non-bloody) stools within a 24-hour period" [9]. In countries where cholera outbreaks have been declared, a suspected cholera case is defined as:

> "Any person presenting with or dying from acute watery diarrhea" [3, 32].

These definitions are fundamental to assessing the success of cholera elimination campaigns such as that of the WHO: a country which has successfully eliminated cholera is defined as one which has:

> "...no confirmed cases [of cholera] with evidence of local transmission for at least three consecutive years and has a well-functioning epidemiologic and laboratory surveillance system able to detect and confirm cases" [9].

1.1.4 – This chapter

In this Introduction, I will first present the mechanisms by which *V. cholerae* is known to cause clinical cholera, as defined above. I will also introduce the genetic determinants that are responsible for this canonical disease, and frame these molecular details in the context of

historical and current cholera pandemics. I will then summarise key microbiological phenotypes that are intimately linked to our understanding of these pandemics, as well as some additional genetic determinants that have been hypothesised to enable certain bacteria to cause pandemic cholera. Finally, I will introduce the understanding gleaned from the application of genome sequencing to studying *V. cholerae*, in order to identify the key aims of this thesis.

1.2 – Vibrio cholerae

1.2.1 – V. cholerae *microbiology*

Since its first description by Pacini in 1854 [33, 34], *V. cholerae* has been recognised as the aetiological agent of cholera. Pacini's original description of *V. cholerae* is considered to be that of the type species of the *Vibrio* genus [35]. During the 1880s, Robert Koch also observed *V. cholerae* in clinical specimens while studying cholera in Egypt, which he termed '*Kommabazillen'*, or '*Vibrio comma*', on the basis of the cellular morphology of the bacterium (reviewed by [36]).

V. cholerae is a comma-shaped Gram-negative bacterium, and typically expresses a monotrichous polar flagellum which can be observed microscopically (Figure 1.1). The species is highly-diverse, exemplified by the fact that more than 200 discrete serogroups of *V. cholerae* have been described [37]. As well as this multitude of antigenic profiles, the species' diversity has also been highlighted by various taxonomic and biochemical studies, which have demonstrated that *V. cholerae* can display numerous phenotypes, ranging from the capacity to utilise certain sole carbon compounds to the ability to secrete haemolysins and other toxins (*e.g.*, [38–41]). These variable phenotypes have been used to assign *V. cholerae* to specific biotypes (section 1.3.1.4). The natural environment for many *V. cholerae* is that of estuarine waters, and can involve the colonisation of chitinous copepods or shellfish (*e.g.*, [42–45]). The species can tolerate a moderate range of salinity and temperature compared to other members of the genus [46], and can metabolise a number of carbon compounds [38] including chitin [47].

Horizontal gene transfer (HGT) is a fundamental aspect to *V. cholerae* genome biology and to the evolution of the species, and will be explored in detail later (section 1.2.5). *V. cholerae* is naturally competent when exposed to appropriate inductive signals or cultured on chitinous

materials [48], harbours a large chromosomal integron (gene capture apparatus) [49], and can employ a type VI secretion system (T6SS) with which to kill adjacent prey bacteria and liberate free DNA into its environment which it can access, using natural competence, to avail of novel genetic material [50, 51]. Bacteriophages, genomic islands, and integrative/conjugative elements also all have roles to play in the evolution of this species [52–56], and will be discussed in subsequent sections.

V. cholerae is an unusual enteric bacterial pathogen because, like other *Vibrio* spp., this species normally has two chromosomes [57]. The chromosome biology of this pathogen is an area of current research, and this species has been used as a model for studying the regulation of chromosomal replication timing and dynamics in bacteria with multiple chromosomes [58, 59]. For example, *V. cholerae* was used to elucidate the role for the *crtS* locus in coupling the replication of chromosome 2 to that of chromosome 1 in *Vibrio* spp., such that chromosome 1 must replicate as far as *crtS* before chromosome 2 replication is initiated [58, 60]. Some rare exceptions have been reported or engineered *in vitro* in which both chromosomes have fused into one macromolecule [61, 62], and in *V. cholerae* in which chromosomes 1 and 2 are fused, the replication origins from both chromosomes may, or may not, be active [63, 64]. It has also been hypothesised that the second *V. cholerae* chromosome, on which the *V. cholerae* integron is located [65], was originally co-opted from a plasmid [66].

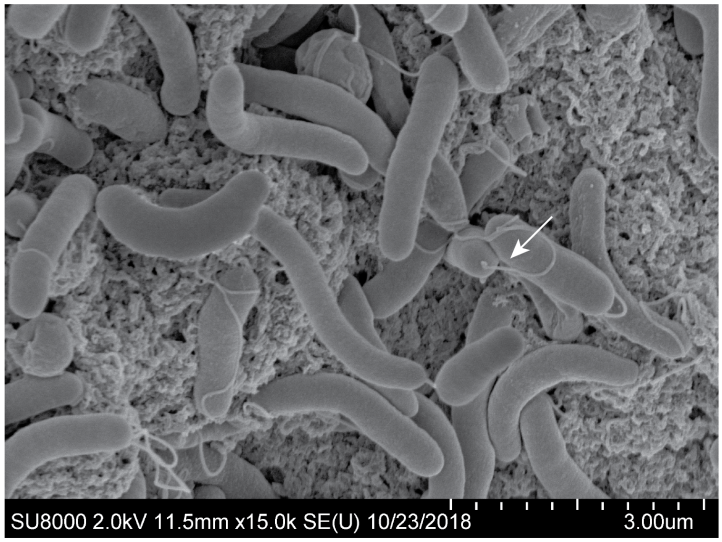

Figure 1.1 – Scanning electron micrograph of *V. cholerae*. SEM produced from fixed *V. cholerae* colonies grown on LB agar plates. Several comma-shaped bacteria can be seen in this image, amongst extracellular polysaccharide (biofilm). A cell producing a polar monotrichous flagellum is indicated (white arrow). Image captured by Claire Cormie.

The specific mechanistic details of how this bacterium causes canonical cases of clinical cholera, and how it acquires the capacity to do this *via* HGT, will now be discussed.

1.2.2 – Mechanism of bacterial pathogenesis in cholera cases

The acute watery diarrhoea that is characteristic of cholera cases is caused by the cholera toxin (CT). This toxin was first discovered as a component produced by *V. cholerae* that was retained in cell-free filtrates produced from bacterial cultures, and produced a strongly-positive reaction in a rabbit ileal loop assay for diarrhoea [67]. CT is an AB_5 exotoxin encoded by the operonic *ctxA* and *ctxB* genes, which are homologous to the *ltA* and *ltB* genes of enterotoxigenic *Escherichia coli* that express the heat-labile enterotoxin, LT [68–70]. CT can be only elaborated by toxigenic *V. cholerae* – that is to say, *V. cholerae* which harbour *ctxAB*. This operon is encoded by a filamentous, lysogenic bacteriophage dubbed CTXφ, the genome of which integrates into the *V. cholerae* chromosome after CTXφ infects the bacterium [52, 71] (section 1.2.3).

CT consists of five identical CtxB subunits and one CtxA protein, and the three-dimensional structure of this hetero-hexamer has been determined [72]. After transcription and translation of *ctxAB*, CT is assembled in the bacterial periplasm [73], from where it is then secreted *via* a type II secretion system (T2SS), as part of the general secretory pathway (Gsp) [74–76]. It has been suggested very recently that specific mutations in the signal peptide sequence of *ctxB* can influence the efficiency by which pre-CtxB is processed and secreted from *V. cholerae*, thereby altering the amount of CT secreted and the relative toxigenicity of an isolate of *V. cholerae* [77].

Once CT is secreted from *V. cholerae*, it is capable of binding to host receptors expressed by the epithelial cells of the small intestine. The primary binding site for CT is the G_{M1} ganglioside [78], which contains galactose and sialic acid residues with which CtxB interacts directly [79]. It should be noted that secondary receptors have been identified to which $CtxB_5$ or the $CtxAB_5$ can bind, including the $Lewis^X$ histo-blood group antigen, L-fucose, and other fucosylated glycoproteins [80–82]. The significance of these secondary receptors in the context of *in vivo* disease is the subject of current research.

Once CT has bound to its receptors, endocytosis occurs, which may take place *via* multiple clathrin-dependent and clathrin-independent pathways [83]. Following endocytosis, CT is trafficked *via* the *trans*-Golgi network and endoplasmic reticulum, where the CtxA protein is unfolded and proteolysed into the A1 and A2 subunits [83–86]. CtxA1 causes the ADP-ribosylation of adenylate cyclase [87–89]. This "locks" adenylate cyclase into a state in which it is bound to GTP, dramatically elevating the rate at which the enzyme converts ATP to cyclic AMP, and increasing the intracellular concentration of cAMP. cAMP-responsive protein kinase (protein kinase A) is stimulated by increased cAMP levels, leading to an activation of the cystic fibrosis transmembrane conductance regulator (CFTR) chloride channel and a loss of chloride ions to the lumen of the intestine [90]. This altered ion gradient means that water is lost from the cell into the intestinal lumen, which is rapidly excreted, forming the 'rice-water' stool that is characteristic of cholera cases. A model of these steps is presented in Figure 1.2.

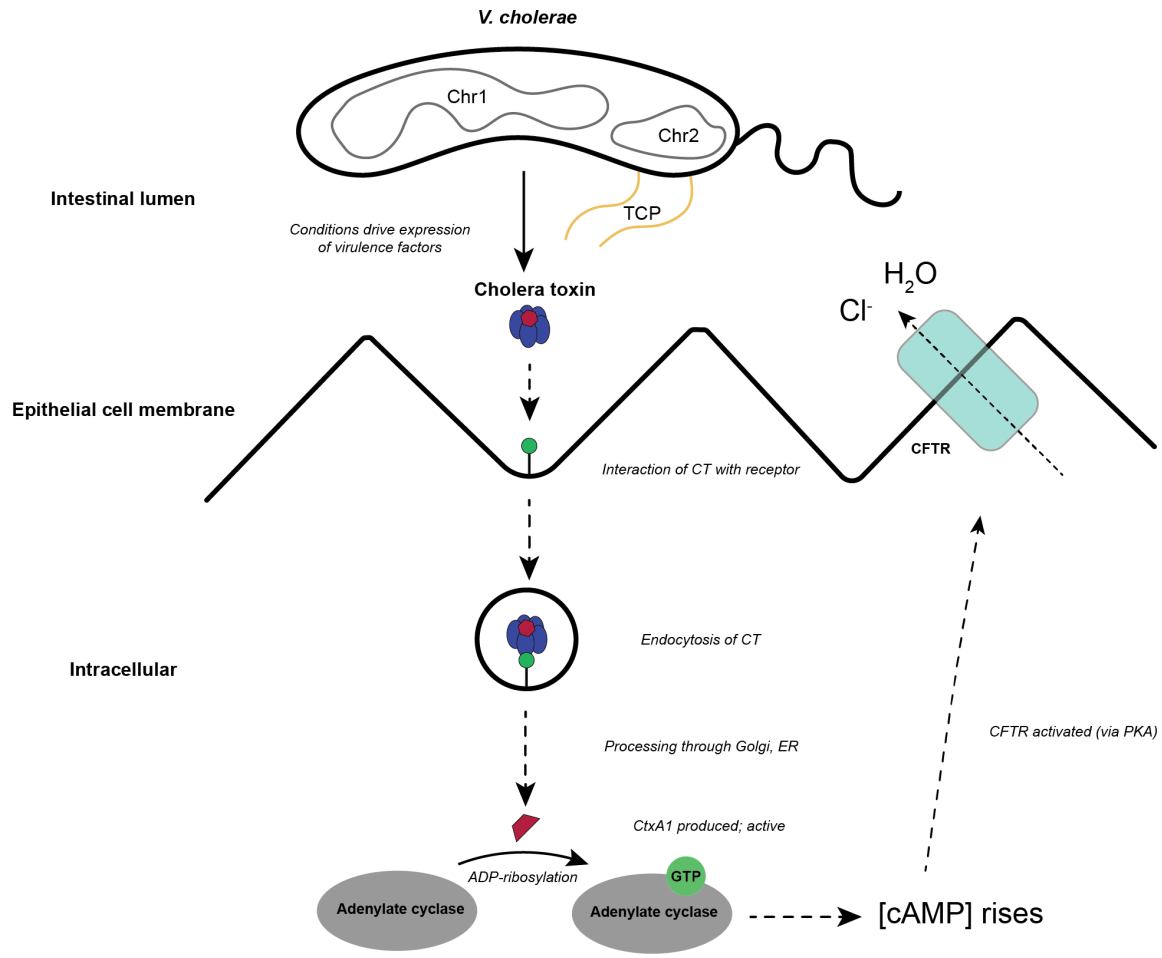

Figure 1.2 – Model of *V. cholerae* pathogenesis leading to the diarrhoea characteristic of cholera. Drawn from steps outlined in section 1.2.2. CT receptors (G_{M1} or others) are indicated as a green circle on the epithelial membrane. TCP is elaborated by *V. cholerae* and acts as a colonisation factor in the intestine. Not to scale.

1.2.3 – Molecular genetics of the CTXφ bacteriophage

The CTXφ bacteriophage infects *V. cholerae via* its receptor, the toxin co-regulated pilus, TCP [55, 91, 92]. Possession of the genes encoding this type IV pilus is therefore canonically necessary for CTX$^-$ *V. cholerae* to be lysogenised by CTXφ, though there is some evidence that other transducing phages can mobilise CTXφ prophages independently of the requirement for TCP [93, 94]. The genes encoding TCP were found to be encoded on a pathogenicity island, initially dubbed the *Vibrio* pathogenicity island (VPI) [55] and now referred to as VPI-1 [54]. VPI-1 contains the *tcp* gene cluster (encoding the TCP receptor) as well as genes encoding an accessory colonisation factor (*acf* cluster) and *toxT*, which encodes a master regulator of virulence gene expression [55] (section 1.2.4).

Upon infection of *V. cholerae*, the CTXφ genome circularises into a replicative form and can then integrate into the bacterial chromosome in an XerCD-catalysed recombination between the CTXφ *attP* site and bacterial *attB* site, producing hybrid *attL* and *attR* sequences [95, 96]. CTXφ typically integrates into the larger *V. cholerae* chromosome, at an *attB* site located near to the *dif* site on chromosome 1. However, it can also occasionally integrate into the smaller chromosome – again, at a site equivalent to the *dif* site on that molecule [97, 98]. The nature of this integration usually sees the integration of CTXφ prophages into the bacterial chromosome in tandem repeats [99]; multiple copies of CTXφ have been suggested to render it impossible for CTXφ to be deleted from certain bacterial strains [69].

CTXφ can replicate itself by producing ssDNA from chromosomal tandem arrays of CTXφ that are integrated in the bacterial chromosome. This is dependent on the product of the CTXφ-encoded *rstA* gene – the RstA protein nicks the CTXφ replication origin located in the Ig-1 intergenic region of the prophage [99, 100]. The exposed 3' site at this nicked site enables synthesis of CTXφ DNA up until the second, tandem CTXφ replication origin is encountered. This second Ig-1 site is also a substrate for RstA cleavage; this second nick creates a free CTXφ genome [99, 100]. Circularisation of this DNA and second-strand synthesis then forms an active replicative phage genome, which can then proceed to be packaged, forming infectious phage particles [101].

1.2.4 – Regulation of ctxAB *and virulence gene expression*

Although CTXφ prophages and their complement of genes are necessary for the production of CT, work in heterologous systems has shown that possession of *ctxAB* alone is not sufficient to express high levels of CT [102]. In order for toxigenic *V. cholerae* to cause the production of rice-water stool, CT must be produced at high levels by the bacterium at an appropriate time for the toxin to bind to receptors on the luminal surface of the human gut epithelium (section 1.2.2). Since *V. cholerae* is a non-invasive pathogen [103], *ctxAB* expression must therefore be activated at an appropriate point in the *V. cholerae* life cycle – *i.e.*, for this to happen, *V. cholerae* must be ingested, survive exposure to environmental stresses in the stomach and upper intestine, and migrate through the mucus layer of the small intestine, to the epithelial cells of the intestine [104]. Consequently, the expression of *ctxAB* and other virulence genes is tightly regulated, the regulatory mechanism of which has been dissected by molecular biologists.

The regulatory decision to activate transcription of *ctxAB* and other *V. cholerae* virulence genes is governed by the activity of the ToxT and ToxR transcription factors. ToxR was first described as a *trans*-factor that positively regulated the synthesis of CT, both in *V. cholerae* and when cloned into *E. coli* harbouring *ctxAB* [102]. ToxR was subsequently shown to be a transmembrane protein with a DNA binding domain [105]. In contrast, ToxT is an AraC-family transcriptional regulator [106] which binds to a specific sequence dubbed the 'toxbox' in the promoters of genes which it regulates [107]. The *toxT* gene is located on VPI-1 and therefore is only present in those isolates which harbour this horizontally-transferred genomic island [55], unlike *toxR*, which is a gene that is found in nearly all *V. cholerae* [108].

The virulence regulon governed by ToxR encompasses the *ctxAB* and *toxT* genes, along with the *tcp* genes required to produce the toxin co-regulated type IV pilus TCP, so named because these genes are co-regulated with *ctxAB* [91]. ToxR also regulates the genes encoding two major *V. cholerae* outer membrane proteins, *ompU* and *ompT* [109]. ToxR can form homodimers with other ToxR proteins, and heterodimers with the ToxS protein [110]. The ToxT protein can activate *ctxAB* and *tcp* promoters independently of ToxRS [109], and also regulates the *acf* genes located on VPI-1 [111]. Expression of *toxT* is directly regulated by ToxRS [112], and TcpP, enhanced by TcpH, also activates *toxT* [113, 114]. TcpH acts by

preventing degradation of TcpP [115]. The *tcpPH* operon, also located on VPI-1, is regulated by other transcription factors such as AphA [116] (see Figure 1.3 for a summary).

Although the cascade which activates CT expression is well-characterised, the integration of extracellular signals that must occur in order to make the 'decision' to activate this cascade is complex, and integrates perception of signals such as pH [117], bile (*via* TcpPH) [118, 119], and the secondary metabolite cyclo(L-phenylalanyl–L-proline) (cyclo(Phe-Pro)), *via* LeuO [120]. The intracellular concentration of cyclic di-nucleotides such as cyclic-di-GMP, modulated by systems such as the VieSAB three-component system, various phosphodiesterases and diguanylate cyclase enzymes [121–124], have been implicated in governing the regulation both of virulence gene expression and of the genes required for *V. cholerae* to switch between environmental and human-virulent behaviours [125, 126]. There is also evidence that global gene expression profiles in *V. cholerae* differ at the early and late stages of infection [127], and that bacteria which have exited a patient in stool have increased fitness and exist in an 'hyper-infectious' state [127–129]. Ablation of the ability of *V. cholerae* to respond to stress, such as by disrupting alternative sigma factors RpoE and RpoS [130, 131], has been shown to attenuate its pathogenicity, and to reduce its ability to reach high titres in the human gut [132]. Taken together, it is clear that the regulation of pathogenicity in *V. cholerae* is complicated, and involves multiple inputs as well as the utilisation of genes encoded both by the core and the accessory genome of the bacterium (summarised in Figure 1.3).

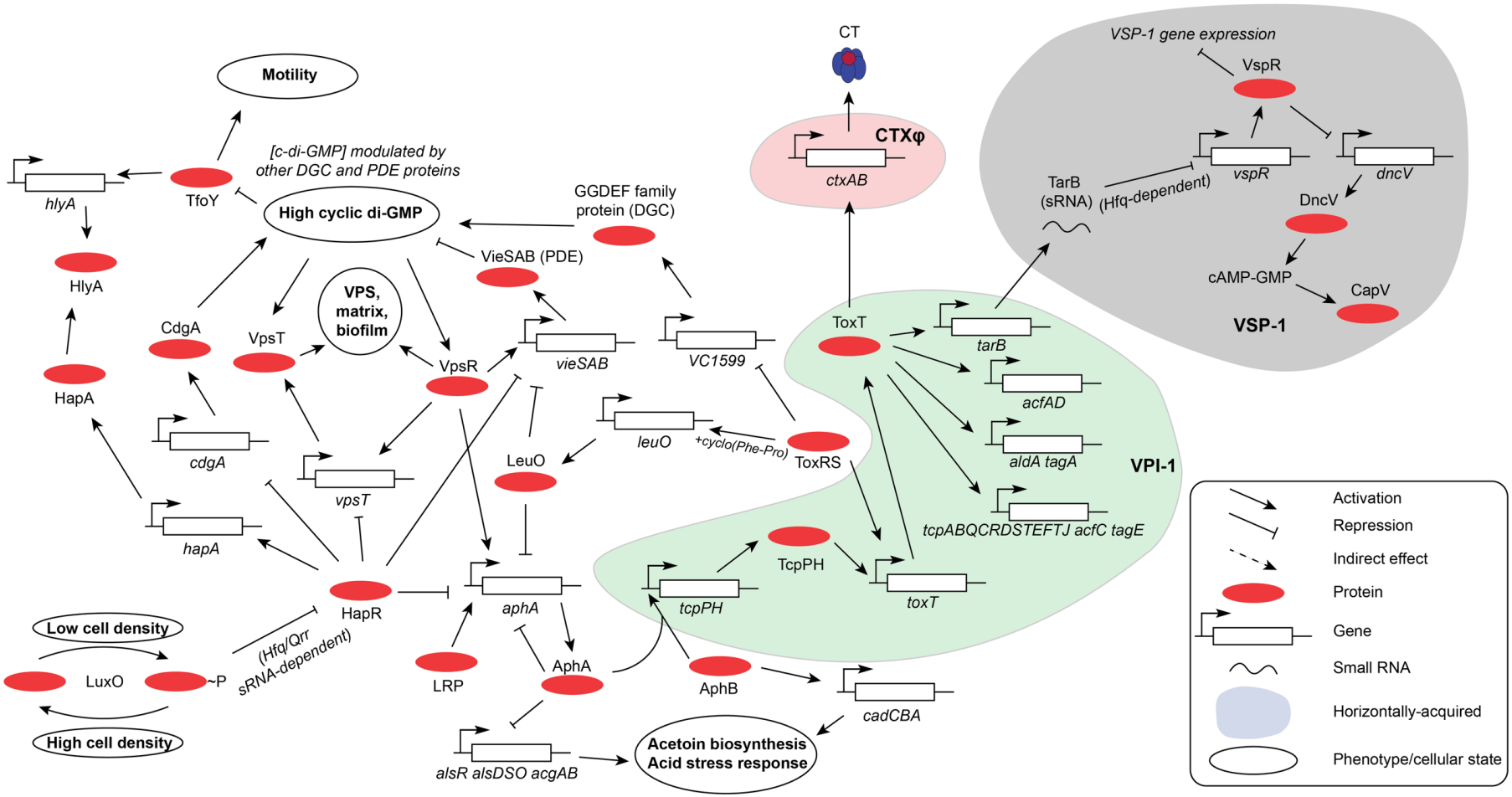

Figure 1.3 – Summary of gene/protein interactions involved in regulating virulence gene expression. This model describes those genes mentioned in this thesis and is not exhaustive. The input of quorum-sensing into this system has been simplified for presentation purposes. Genes that are encoded on mobile genomic islands or are otherwise horizontally-acquired are enclosed in coloured sub-sections of the figure; all other genes are part of the core *V. cholerae* genome. The specific details of AphA/AphB regulation of gene expression are discussed in Chapter 5.

1.2.5 – Importance of horizontal gene transfer in V. cholerae *biology and pathogenicity*

As mentioned previously (section 1.2.3), the principal *V. cholerae* virulence determinants are encoded on mobile genetic elements, either CTXφ or VPI-1. These are mobile genetic elements that either actively engage in HGT (CTXφ) or engaged in this process ancestrally (VPI-1). Genomic islands other than VPI-1 also play important roles in *V. cholerae* biology. Following the identification of genes specific to classical and El Tor *V. cholerae* by Dziejman *et al*. [133], VPI-2 was formally characterised [134]. This genomic island encodes *V. cholerae* neuraminidase (sialidase, now NanH; EC 3.2.1.18), an enzyme which was first described in 1947 as a protein which destroyed the receptor for influenza virus – hence, the protein was first known as the 'receptor destroying enzyme' [135]. The *nanH* gene was first cloned in 1988 [136], and its product has been shown to be capable of hydrolysing higher-order gangliosides to that of G_{M1}, [137]. However, purified sialidase does not appear to change the profile of toxin binding sites in the intestine of multiple species, *in vitro* or *in vivo* [78]. It remains uncertain whether this hydrolysis is to increase the number of potential receptors for CT during an infection or for growth, since it is known that *V. cholerae* can utilise sialic acid as a sole carbon source [138].

Other *V. cholerae* genomic islands and prophages have also been identified, and their distribution amongst limited numbers of diverse *V. cholerae* have been described (*e.g.*, [54, 139–141]). Two important genomic islands, VSP-1 and VSP-2, have been identified from microarray data and comparative genome sequencing which have been purported to be found exclusively in those *V. cholerae* that cause current cholera pandemics [54, 133]. The function of these is not completely understood, though some work has been done on VSP-1 function [142]. Small RNAs (sRNAs) regulated by ToxT, such as TarB, modulate the expression of genes within VSP-1, including those genes encoding dinucleotide cyclase, which produces cyclic AMP-GMP (c-AMP-GMP) [142] (Figure 1.3). These cyclic dinucleotides and VSP-1 have been implicated in modulating the ability of *V. cholerae* to colonise the intestine [142]. The function of VSP-2, and its variants, remains uncertain [143].

Recombination has been shown to be important to *V. cholerae* evolution. For instance, recombination within the chromosomal region encoding the O-antigen has been postulated to be a mechanism by which *V. cholerae* may have undergone serogroup conversion [144, 145] (discussed further in section 1.3.2). Conversion of serotype is of particular clinical significance

because of the important immunogenic properties of specific O-antigens in cholera vaccine efficacy [146] (section 1.3.1.3).

The evolution in recent years of many bacterial pathogens has been driven in part by the increased use of antimicrobials, and the emergence of antimicrobial resistant pathogens is a consequence of the selective pressures imposed by these drugs [147]. There is evidence of *V. cholerae* O1 acquiring multidrug resistance plasmids [148, 149] and resistance genes within the chromosomal integron [150–152], including resistance to fluoroquinolones both by acquisition of *qnr* genes [150, 153, 154] and by mutation in *parC* and *gyrA* [155]. Resistance determinants can also be encoded by a conjugative ICE-type element, dubbed SXT, which was first described as an element conferring resistance to multiple antibiotics in *V. cholerae* O139 [156]. The SXT element is conjugative and self-transmissible, integrating into the *V. cholerae* chromosome in a RecA-dependent manner specifically within the *VC_0659* locus [54, 59, 156].

Unusually for a pathogenic bacterium, *V. cholerae* evolution does not appear to be driven by the need to acquire antimicrobial resistance (AMR) in response to direct therapeutic antibiotic usage; it should be noted that the antimicrobials to which these determinants render *V. cholerae* resistant (*e.g.*, sulfamethoxazole, trimethoprim and streptomycin as in the case of SXT [156]) are not drugs which are used to treat cholera (section 1.1.2). There are, however, some examples of drug resistance being acquired by *V. cholerae* in response to therapeutic use of antimicrobials – for instance, as a consequence of the use of tetracycline to control a sensitive strain in Tanzania, *V. cholerae* O1 became resistant to tetracyclines upon the acquisition of an IncA/C resistance plasmid [157, 158]. It should also be emphasised that our understanding of antimicrobial resistance amongst *V. cholerae* is largely limited to the study of pandemic *V. cholerae* lineages, as the aetiological agents of pandemic cholera; we know much less about the distribution of resistance determinants amongst more diverse or non-pathogenic members of the species. This is a topic which will be addressed in this thesis (see Aims, section 1.5).

1.3 – Pandemic and non-pandemic cholera

1.3.1 – Pandemic cholera

The discussion of Asiatic cholera above (sections 1.1, 1.2) describes a disease which is canonically caused by specific clones of *V. cholerae* serogroup O1 [1, 36], and it is dependent

on the activity of CT on epithelial cells of the human intestine. However, when considering pandemic cholera, it is useful to recall that a pandemic of a disease, such as influenza, can be defined as:

> "an epidemic occurring worldwide, or over a very wide area, crossing international boundaries and usually affecting a large number of people" [159, 160].

It is therefore important to bear in mind that for a clone of *V. cholerae* to be both the aetiological agent of cholera and of pandemic cholera, it must be capable of causing the disease described in section 1.1.1, as well as spreading rapidly across the globe. This point will be emphasised by several historical epidemics of cholera and cholera-like disease that were not caused by pandemic *V. cholerae* (sections 1.3.2 and 1.4.3).

1.3.1.1 – History of cholera pandemics

Seven pandemics of cholera have been described in recorded history [36, 161], the first six of which are believed to have been caused by *V. cholerae* of the classical biotype (a set of phenotypic tests used to identify certain *V. cholerae* - details of biotyping will be presented in section 1.3.1.4). It is important to state that although there is direct evidence that the sixth, fifth, and second pandemics were caused by classical biotype *V. cholerae* [97, 162, 163], it is inferred that the remaining historical pandemics were caused by classical *V. cholerae*. Snow's work, carried out during the second pandemic [30], would therefore have described cholera caused by classical biotype *V. cholerae*. Hereafter, the term 'classical' is used to describe the biotype, and 'Classical' is used to denote the phylogenetic lineage comprised of classical biotype *V. cholerae* (see section 1.4).

The twentieth century saw the transition from the sixth to the seventh cholera pandemic. The sixth pandemic occurred between 1899 and 1923 [36, 97, 163], and the seventh cholera pandemic began in 1961 [36, 164]. The onset of the seventh pandemic caused alarm because the *V. cholerae* which caused this pandemic were of the El Tor biotype rather than the classical biotype [161, 164, 165]. Consequently, and due in part to confusion in the nomenclature in use at the time, these El Tor biotype *V. cholerae* were referred to as "*Vibrio paracholera*" [165, 166] (discussed further in section 1.3.2).

A detailed recapitulation of the history of the seventh cholera pandemic is beyond the scope of this thesis, but has been presented by several groups [25, 161, 164, 167–169]. Briefly, the seventh pandemic began in Sulawesi, Indonesia, in 1961 [161] and spread into Southeast Asia [165, 170], from where it subsequently spread to Africa by the early 1970s [158]. Cholera spread within the continent between 1970 and the early 1990s (reviewed by [167, 171, 172]), prior to its transmission into Latin America in the early 1990s. In 1991, cholera broke out in Lima, Peru, following a period of nearly 100 years in which South America was free of cholera epidemics [168, 173]. This epidemic proceeded to spread to other countries within Central and South America [174], though the highest case/fatality rates were associated with countries in Central America [175].

More recently, cholera epidemics have been the subject of intense coverage by the popular press. In October 2010, an outbreak of cholera in Haiti began [176], which led to over 170,000 infections and 3,600 deaths by December 2010 [177, 178]. Cholera cases continued to be reported between 2010 and 2017 [179]. The cholera epidemic in Yemen, which began in late 2016 and continues today, was a similarly high-profile event [28, 180] and led to considerable numbers of disease cases – by 12th March 2018, 1,103,683 suspected cholera cases and 2,385 deaths from cholera had been reported in Yemen [28]. These statistics underline the fact that pandemic cholera remains a serious and current public health concern.

1.3.1.2 – The cholera paradigm

As mentioned earlier (section 1.2.1), *V. cholerae* is capable of living in estuarine and brackish water, often in association with copepods and crustaceans with chitinous exoskeletons [45, 181–183]. It has been observed repeatedly that cholera outbreaks are correlated with seasonality and rainfall (*e.g.*, [184]). It has also been reported that *V. cholerae*, and other bacteria, can enter a 'viable but non-culturable' state of dormancy [185], in which bacteria were shown to be metabolically active (viable) but could not be cultured using standard microbiological methods [185, 186].

Taken together, these observations led to the hypothesis that *V. cholerae* is autochthonous to estuarine environments [29, 187], and to the proposition of the 'cholera paradigm' [29]. In practical terms, this model assumes that environmental reservoirs harbour local populations of *V. cholerae* in an estuarine environment in which the bacterium is known to live [188]. Upon

exposure to favourable environmental or climactic conditions, including temperature, salinity, *etc*., these bacterial populations expand in size and can be ingested by humans, causing outbreaks of cholera [29]. This model suggests that cholera outbreaks result from the expansion of local populations of *V. cholerae* that reside within an area or environment which are then ingested by humans, and that climactic factors are the principal driving force behind cholera outbreaks (rather than human-to-human transmission of *V. cholerae*).

Many of the data upon which the cholera paradigm was formed were carried out in the Bay of Bengal, and their applicability to other environments (particularly inland areas) which experience cholera epidemics and hotspots has been contested [167]. Genomic analysis of the *V. cholerae* lineage causing the current cholera pandemic also indicates that the cholera paradigm, and its reliance on local populations of *V. cholerae* seeding cholera outbreaks, is not consistent with the observation that a single lineage of toxigenic *V. cholerae* has caused pandemic cholera since 1961 (section 1.4.2; [158, 189]).

1.3.1.3 – V. cholerae *serogroups and serotypes*

As mentioned throughout section 1.3.1, cholera pandemics are caused by *V. cholerae* of serogroup O1. Nonetheless, over 200 serogroups of *V. cholerae* have been described on the basis of variation in the O-antigen of the bacterial lipopolysaccharide (LPS) [37, 190]. The *V. cholerae* LPS is highly immunogenic, and antibodies against LPS have been shown to mediate near-exclusive immunity to *V. cholerae* in both humans and animals [103]. In rabbit immunisation experiments, a highly synergistic immunity was conferred by purified LPS (serogroup O1) and CT when simultaneously administered to animals [103, 191]. Although all pandemic cholera to date has been caused by *V. cholerae* O1, it is important to state that some large-scale outbreaks in Southeast Asia have also been caused by *V. cholerae* O139; the details of these outbreaks and how they are related to pandemic *V. cholerae* O1 will be discussed in section 1.3.2.1.

As early as the 1930s, when Gardner and Venkatraman studied the "original, varied, and middle" phenotypes of *V. cholerae* agglutination to specific O1 sera, it was recognised that there was additional subtlety to the serogrouping of pandemic *V. cholerae* O1 [192]. This was subsequently found to be due to Inaba/Ogawa serotype variation. "Inaba" and "Ogawa" were the names of two cholera patients in 1921 from whom the strains used to describe these

serotype variants were obtained [193]. A rare third serotype variant, Hikojima, has also been described – this is an unstable mixed phenotype in which an isolate simultaneously expresses Ogawa and Inaba antigens, though Hikojima isolates will ultimately type as Inaba [194–196].

Ogawa serotypes are a result of wild-type activity of the WbeT protein (formerly named RfbT [197, 198]), which methylates the terminal perosamine sugar on the O1 lipopolysaccharide chain [197–199]. In the absence of this methyl group, an Inaba phenotype results, and *V. cholerae* will be agglutinated by Inaba rather than Ogawa antisera [197–199]. Mutations in *wbeT* that lead to an abolition of WbeT activity result in seroconversion of *V. cholerae* O1 from Ogawa to Inaba serotype [197–200]. There is evidence that Ogawa to Inaba mutations occur frequently amongst *V. cholerae*, and also that reversion from Inaba to Ogawa serotype can occur *in vivo*, albeit rarely [197, 201, 202].

Inaba and Ogawa serotypes are significant because both elicit different immunological responses [20, 203]. Thus, they are both included in the formulation of cholera vaccines, such as Dukoral™ [204]. Co-expression of the Inaba and Ogawa antigens by a stable Hikojima strain has also been exploited for vaccinology purposes [196, 205]. Serotyping of *V. cholerae* O1 continues to be a clinically relevant microbiological test performed on bacterial isolates [3, 32], and diagnostic laboratories [206] and epidemiologists [207] also serotype *V. cholerae* O1 as a matter of routine. Outbreaks of cholera are often described in terms of the serotype of *V. cholerae* O1 that is associated with the outbreak – for instance, the initial cholera epidemic in Peru, 1991 was associated with Inaba isolates and with Ogawa isolates in subsequent years [208]. Thus, both the serogroup and serotype of *V. cholerae* have historically been important phenotypes for understanding cholera epidemiology. This point will be re-visited later in this thesis (section 3.4.6).

1.3.1.4 – Classical and El Tor biotypes

The first six cholera pandemics were caused by serogroup O1 *V. cholerae* of the classical biotype (section 1.3.1.1). Thus, toxigenic *V. cholerae* isolated during the sixth pandemic were those used to establish the biochemical, taxonomic, and microbiological criteria needed to classify a bacterium as "*V. cholerae*" and the aetiological agent of cholera [209]. In contrast, the seventh cholera pandemic, which began in 1961, is caused by an El Tor biotype *V. cholerae* O1 [164]. The taxonomic relationship between El Tor and classical biotype *V. cholerae* has

been disputed, and it had been proposed that both comprised separate species [36, 164]. This was subsequently overturned, and both biotypes were re-classified as distinct members of the same species on the basis of their microbiological and biochemical properties [164, 210].

In 1905, Gotschlich made the first report of *V. cholerae* which displayed a different biochemical phenotype to the strains now referred to as being of the classical biotype [211]. These unusual *V. cholerae* were isolated from patients at the El Tor quarantine camp in Egypt leading to bacteria displaying this phenotype being dubbed 'El Tor biotype' *V. cholerae*. There are also reports from the early twentieth century of "El Tor" vibrios distinct from the "vibrio of cholera" having been isolated from patients suffering from dysentery and colitis [212]. The term "paracholera" was used to describe cases of disease associated with these El Tor vibrios, not least because patients suffering from cholera *sensu stricto* (caused by classical *V. cholerae*) were subject to quarantine [213]. Paracholera was a term used historically to describe cases of disease which resembled cholera, but were not caused by *V. cholerae* as described by Koch (*e.g.*, [214, 215]). However, there were inconsistencies in the bacteriological reports describing infections giving rise to cholera and cholera-like disease [216]; as noted by Mackie:

> "The paracholera vibrios comprise a group which is not serologically homogenous, but ... represents a considerable number of serological races..." [217].

The importance of the ability to discriminate between *V. cholerae* associated with pandemic cholera, and those causing sporadic disease, was also recognised in historical reports. de Moor, in 1949, cites and translates a quote from as early as 1913 [218]:

> "with the same necessity with which paratyphoid is distinguished from typhoid, an 'El Tor disease' should be distinguished from true cholera" [218].

These observations have also been recapitulated in recent years – Salim and colleagues noted in 2005 that many environmentally-isolated *V. cholerae* displayed El Tor phenotypes, stating:

> "The properties that characterise the El Tor biotype are those of environmental strains" [163].

Although "El Tor" has come to be synonymous with the pathogen of the seventh pandemic [2, 164], the above quotations and discussion illustrates the fact that the phenotypes associated with El Tor *V. cholerae* both describe the bacterium which causes current pandemic cholera as well as other *V. cholerae*, which may cause sporadic infections or be non-pathogenic. *V. cholerae* are biotyped on the basis of a set of biochemical and microbiological tests, detailed below (Table 1.1).

	Biotype	
Test	***Classical***	***El Tor***
Haemolysis	Negative	Positive
Voges-Proskauer test	Negative	Positive
Haemagglutination (chick or sheep erythrocytes)	Negative	Positive
Polymyxin B, 50 units	Susceptible	Resistant
Classical phage IV	Susceptible	Resistant
El Tor phage 5	Resistant	Susceptible

Table 1.1 – Summary of *V. cholerae* O1 biotyping phenotypes. Scheme modified from [40, 206].

The molecular basis of each of these phenotypes will be discussed in detail in Chapter 5. However, it is useful to note at this juncture that microbiologists have sought to explain the molecular basis of the variation in these phenotypes amongst strains. For instance, haemolysis in *V. cholerae* is mediated by the HlyA haemolysin, encoded by *hlyA*. This gene is truncated in strains of classical biotype *V. cholerae*, such that the C-terminal domain of HlyA cannot be produced [219]. Although the N-terminal domain of HlyA can be produced by classical *V. cholerae* and is cytotoxic to mammalian cells, the C-terminal domain has been shown to be necessary for haemolysis [219, 220]. Recently, pandemic *V. cholerae* have been isolated which exhibit "hybrid" biotypes – *i.e.*, a mixture of classical and El Tor phenotypes, including a loss of haemolysis [221, 222]. However, sequencing of the *hlyA* locus has demonstrated that these hybrid phenotypes can be due to multiple independent *hlyA* mutations, not due to the acquisition of the same *hlyA* mutation as found in classical isolates [221].

1.3.2 – Non-pandemic cholera

A fundamental issue in the study of cholera and *V. cholerae* has been expressed succinctly by Kaper *et al*.:

> "*V. cholerae* is a well-defined species on the basis of biochemical tests and DNA homology studies. However, this species is not homogeneous with regard to pathogenic potential." [1].

Although cholera pandemics are associated with serogroup O1 *V. cholerae* (section 1.3.1), there are historical examples of outbreaks and epidemics caused by non-O1 *V. cholerae*. The most well-described of these are the epidemics caused by *V. cholerae* of serogroups O139 and O37.

1.3.2.1 – V. cholerae *O139*

In 1992, the dogma that *V. cholerae* O1 was the exclusive agent of epidemic cholera was challenged by the sudden occurrence of cholera epidemics in South Asia caused by serogroup O139 *V. cholerae*, which caused a large cholera outbreak across Bangladesh and India [223–225]. The substantial numbers of cholera cases caused by *V. cholerae* O139 in Southeast Asia during the early 1990s led to fears that this serotype would emerge as the aetiological agent of an eighth cholera pandemic [226–228]. However, these fears were ultimately unfounded. After the initial 1992-93 epidemic, *V. cholerae* O139 was only associated with low numbers of cholera cases, and did not proceed to cause a global pandemic. However, a second outbreak associated with *V. cholerae* O139 occurred in Bangladesh during early 2002 [229]. The re-emergence of this serogroup renewed fears of an eighth cholera pandemic driven by *V. cholerae* O139 [230], but once again, this clone did not proceed to cause a cholera pandemic.

The disease caused by *V. cholerae* O139 is clinically indistinguishable from that caused by *V. cholerae* O1 [224]. Molecular evidence also indicated that *V. cholerae* O139 was closely related to epidemic *V. cholerae* O1 [231], and data were subsequently obtained that showed that natural competence and homologous recombination enabled the *in vitro* exchange of the O1 and O139 operons, suggesting that such an event had occurred to seroconvert pandemic *V. cholerae* O1 to serogroup O139 [144]. Early genetic and biochemical studies demonstrated that O139 strains were closely related to O1 seventh pandemic El Tor strains, and it was suggested that *V. cholerae* O139 had arisen from an O1 El Tor ancestor [223, 231–233]. Whole-genome sequencing later confirmed this hypothesis, showing that toxigenic *V. cholerae* O139 were closely related to the *V. cholerae* O1 El Tor clone causing the seventh cholera pandemic [54, 234, 235]. This clone is described more fully in section 1.4.2.

As well as serogroup, there are other notable differences between *V. cholerae* O139 and *V. cholerae* O1. For instance, *V. cholerae* O139 expresses a polysaccharide capsule, which *V. cholerae* O1 isolates do not [236]. The capsule is encoded by genes absent from the *V. cholerae* O1 causing current pandemic cholera, and these genes are located adjacent to the locus encoding LPS biosynthesis genes in *V. cholerae* O139 [145, 230, 237–240]. The complement of genomic islands in *V. cholerae* O139 is also different to that of pandemic *V. cholerae* O1, an observation first made using the genome sequence of MO10, a *V. cholerae* O139 isolated in India during 1992 [54, 133, 241].

V. cholerae O139 remains an important organism to study and to monitor, not least because this serogroup has continued to be isolated since 2002. Recently, non-toxigenic *V. cholerae* O139 have been isolated in Thailand [242]. Toxigenic *V. cholerae* O139 have been isolated in China as recently as 2013 [243], and continue to be isolated in Bangladesh [244]. Accordingly, *V. cholerae* O139 continues to be the subject of surveillance in Southeast Asia. Crucially, it has been demonstrated in animal models that immunisation with serogroup O1 vaccine strains of *V. cholerae* does not confer cross-protection against infection with *V. cholerae* O139 [146]. Both serogroups therefore continue to be included in killed whole-cell vaccines [245–247], including Shanchol™, Euvichol™, mORC-Vax™, and Cholvax™ [20].

1.3.2.2 – V. cholerae *O37*

Another key example of epidemics caused by non-O1 *V. cholerae* are the outbreaks caused by *V. cholerae* O37. In November 1968, a severe outbreak of gastroenteritis occurred in Idd Eltin, Kassala Province, Sudan [248]. The outbreak was associated with a newly-opened well, around which tens of thousands of people were reported to have gathered without sanitation provisions [248]. Cholera was suspected, and non-agglutinable Heiberg group I Vibrios (metabolising mannose and saccharose, but not arabinose [40, 249], which is a definition now known to encompass pandemic *V. cholerae* O1 [41]) were isolated from stool and rectal swab samples [248]. During 1965, a similar outbreak of gastroenteritis occurred in Czechoslovakia amongst individuals at an automobile training centre [250]. Patients were described as producing stool that contained neither blood nor mucus [250]. A vibrio was isolated from these specimens which was not agglutinated by *V. cholerae* Inaba or Ogawa sera [250]. Subsequently, an isolate from this outbreak, “280 NAG” [251], was deposited in the American Type Culture Collection

under accession number ATCC 25872. The ATCC metadata and its initial report for this strain lists it as having been isolated from a 'patient with clinical cholera' [251, 252].

Isolates from the Sudanese and Czechoslovakian outbreaks have been shown to be *V. cholerae* of serogroup O37 [144, 253]. The O37 serogroup was first defined in 1970 [254], the type strain for which was isolated in India in 1969 [190, 255, 256]. ATCC 25872 was used in the original characterisation of VPI-1, and is both CTXφ and VPI-1 positive [55], though different *tcpA* alleles have been shown to be harboured by VPI-1 in various *V. cholerae* O37 [257]. Allelic variants of *ctxB* have similarly been reported amongst toxigenic *V. cholerae* O37 [189]. It has been demonstrated that genomic DNA prepared from O37 serogroup strain ATCC 25872 could transform naturally-competent *V. cholerae* O1 and convert them to serogroup O37, just as was shown for *V. cholerae* O139 (section 1.3.2.1) [144]. ATCC 25872 and other *V. cholerae* O37 have also been shown to have a constitutively-active T6SS, making it a useful strain for the study of intra- and inter-strain competition [258].

Several studies have found that these toxigenic *V. cholerae* O37 are closely-related to pandemic *V. cholerae* O1 [255, 259–265] and this has been supported by whole-genome sequencing data [54, 189, 234]. However, Bik *et al* noted that *V. cholerae* O37 'exemplifies the pitfalls of using phenotypic methods to discriminate *V. cholerae* strains' [259]. Although multiple clinical and environmental isolates of serogroup O37 *V. cholerae* have proven to be both toxigenic and phylogenetically-related to O37 isolates from the Sudanese and Czechoslovakian outbreaks and, therefore, to pandemic *V. cholerae* [253], other *V. cholerae* O37 have been isolated which are distantly related, or are non-toxigenic and lack VPI-1 [259–261, 266–268].

1.3.2.3 – Cholera on the Gulf Coast

During the 1970s and 1980s, toxigenic *V. cholerae* O1 El Tor were isolated from cases of cholera in Texas and Louisiana, USA [269, 270]. Non-O1 *V. cholerae* were also recovered from patients with diarrhoea in Louisiana, which were toxigenic and produced detectable CT [269]. In 1981, a cholera outbreak occurred on an oil rig south of Port Arthur, Texas, USA [271]. Toxigenic, haemolytic *V. cholerae* O1 Inaba was isolated from the stool of the index patient [271]. These outbreaks led to the hypothesis being proposed that *V. cholerae* is endemic on the US Gulf Coast [272], though it has also been recognised that the low frequency at which this clone can be isolated from the environment means that the relative contribution of human-

to-human transmission and environmental reservoirs in the dynamics of this Gulf Coast clone remain unclear [269].

Molecular data have shown that this Gulf Coast clone is distinct from the classical and El Tor biotype pandemic clones [273, 274]. Isolates from this outbreak were first sequenced in 2009, which were shown to be related to the classical and El Tor biotype pandemic clones [54]. This phylogenetic positioning was corroborated in 2011, using both toxigenic and non-toxigenic *V. cholerae* O1 isolates from the Gulf Coast [234]. These isolates harbour VPI-1 and VPI-2, but not VSP-1 or VSP-2 [189].

It should be noted that as well as these *V. cholerae* O1, isolates of *V. cholerae* O75 have also been reported to cause sporadic cases and outbreaks of cholera in the vicinity of the Gulf Coast in the USA [275, 276]. These *V. cholerae* O75 were toxigenic, and harboured VPI-1 (with the classical *tcpA* variant), a VPI-2 variant harbouring a T3SS, and a VSP-2 variant [276].

1.3.2.4 – Non-O1/O139 V. cholerae *infections and virulence determinants*

Thus far, consideration has been given to the aetiological agents of current and historical pandemics (section 1.3.1) and to bacteria that have caused epidemics of cholera but did not proceed to cause global pandemic disease (sections 1.3.2.1 – 1.3.2.3). All of the *V. cholerae* considered up until now have been toxigenic, causing disease by virtue of expressing CT and inducing choleraic diarrhoea (section 1.2.2). However, although toxigenic *V. cholerae* is notorious for causing cholera epidemics and pandemics, sporadic cases of disease can also be caused by this species. Approximately 40 cases of disease caused by non-O1/O139 *V. cholerae* are reported annually to the CDC [277]. These non-O1/O139 *V. cholerae* may cause gastroenteritis, extraintestinal infections (such as wound and skin infections) or septicaemia [277–282], and have been identified as causing sporadic outbreaks in studies that focus on epidemic cholera (*e.g.*, [189, 235, 283]). Crucially, these non-O1/O139 *V. cholerae* have never been observed to proceed to cause epidemic cholera. It is important to state that there are rare examples of non-O1/O139 *V. cholerae* being toxigenic and harbouring CTXφ [54, 189, 276].

Several accessory virulence determinants can be expressed by *V. cholerae*, the most important of which will be discussed below. Some of these are encoded as part of the *V. cholerae* accessory genome, either on genomic islands or on bacteriophages. Others are part of the core

genome and are common to nearly all *V. cholerae* which have been characterised. For instance, possession of TCP, and the VPI-1 genomic island which encodes it, is necessary to render a strain of *V. cholerae* susceptible to infection by CTXφ, the generalised transduction of CTXφ by other bacteriophages notwithstanding [93]. However, TCP is itself a virulence determinant, and the TcpA pilus acts as a colonisation factor, enabling *V. cholerae* to adhere to the intestinal epithelium [284]. Other elements act as virulence determinants, such as the accessory colonisation factor Acf (also encoded by VPI-1), and play important roles in *V. cholerae* colonisation and chemotaxis [111, 285, 286].

The heat-stable enterotoxin of non-agglutinable *V. cholerae*, NAG-ST, was first observed in the early 1980s [263, 287] and the gene encoding this toxin was sequenced in 1990 [288]. NAG-ST is similar to the heat-stable toxin produced by enterotoxigenic *E. coli* [288, 289], and *V. cholerae* encoding this virulence determinant have been associated with causing severe diarrhoea in volunteer studies [290]. Strains of non-O1 *V. cholerae* encoding this toxin have been isolated from environmental sources and seafood farms [291, 292], and there are some reports of *V. cholerae* O1 encoding NAG-ST [293, 294]. Other Vibrios, including *Vibrio mimicus*, have been shown to encode NAG-ST [295].

V. cholerae can produce a multifunctional autoprocessing RTX (MARTX) toxin, encoded by a gene cluster adjacent to the CTXφ integration site on chromosome 1 [59, 296, 297]. This toxin is cytotoxic, and can affect cytoskeletal structure in target eukaryotic cells by interfering with actin crosslinking [298, 299]. Thus, this toxin is believed to contribute to the diarrhoeal and inflammatory symptoms occasionally experienced by human subjects when exposed to *V. cholerae* vaccine strains [297] and to lethality in mouse models of infection [300]. The *rtx* gene cluster typically contains two divergently-transcribed operons – *rtxHCA*, where *rtxA* is a long gene encoding the toxin (the longest gene in the *V. cholerae* genome [59]), and the *rtxBDE* operon which encodes a type I secretion system (T1SS) [299, 301]. RtxA is exported from the bacterial cytosol *via* this T1SS [299]. The RtxA toxin is produced by nearly all clinically- and environmentally-isolated *V. cholerae* [297, 299], though certain environmental isolates may encode RtxA variants [302], and a deletion of *rtxC* and the 5' sequence of *rtxA* has been described in classical isolates [297].

Certain *V. cholerae* harbour gene clusters that encode type III secretion systems (T3SS) integrated into the same chromosomal location as VPI-2 [141]. A T3SS is a macromolecular

apparatus capable of translocating and injecting potentially-cytotoxic effector proteins from *V. cholerae* into adjacent prokaryotic and eukaryotic cells [303], unlike the mechanism of action of CT (section 1.2.2). These systems are common virulence determinants in other *Vibrio*, such as *V. parahaemolyticus* [303], but are present in certain non-O1/O139 *V. cholerae* [141]. There have been reports of serogroup O1 *V. cholerae* harbouring T3SS genes, but these have not yet been characterised in detail [304]. These systems are linked to causing rapid-onset inflammatory diarrhoea caused by CTXφ-negative *V. cholerae* in the infant rabbit model [305]. The first *V. cholerae* strain harbouring a T3SS to be characterised in detail was AM-19226, which was shown to encode a T3SS similar to that of *V. parahaemolyticus* [266]. It has been noted that T3SS in *V. cholerae* and *V. parahaemolyticus* are more similar to one another than the two species are to one another, strongly suggesting that these systems can be exchanged amongst *Vibrio* species [303].

1.4 – Insights from V. cholerae *genomics*

1.4.1 – Comparative V. cholerae *genomics*

The comparative genomics of *V. cholerae* have been studied using several approaches in the past. For example, microarray hybridisation technology has been used to describe variations in gene content amongst small numbers of pandemic and non-pandemic *V. cholerae*. The *V. cholerae* microarray, designed using the N16961 reference sequence, has been used for this comparative work [133]. However, microarray hybridisation cannot identify or describe genetic novelty – the approach can only detect the presence and absence of genes that were in the genome sequence used to generate the array [306]. Comparative genomic approaches have also been used to analyse the genome sequences of key fully-sequenced strains of *V. cholerae*, including O395, an isolate representing the classical biotype, pre-seventh pandemic reference sequences M66 [97, 307] (see section 1.4.2), and other isolates of particular importance to researchers (*e.g.*, [98]). Whole-genome sequencing of *V. cholerae* has allowed for comparative genomics analyses to be performed, yielding similar results to these microarray experiments [54]. However, most of these studies implement short-read sequencing (*e.g.*, Illumina) with read lengths of 150 bp, up to a maximum of 300 bp.

1.4.2 – V. cholerae *genomics and genomic epidemiology*

Prior to the implementation of whole-genome sequencing, the phylogenetic analysis of housekeeping genes had indicated that the classical biotype *V. cholerae* clone and the El Tor clone causing the seventh cholera pandemic were distinct from one another [163]. In that study, the relationship between M66, and the classical and El Tor epidemic clones, was also investigated (M66 was isolated in Makassar, Indonesia in 1937). Consistent with its time and place of origin (section 1.3.1.1), M66 was shown to represent a "pre-pandemic" ancestor of the current El Tor pandemic clone [163, 213, 234]). The relationships of other El Tor clones to the seventh pandemic clone was also described [163].

The *V. cholerae* reference genome was published in 2000 [59]. This reference sequence was produced from N16961, a toxigenic, serogroup O1 Inaba, biotype El Tor *V. cholerae* isolated from a patient in Bangladesh in 1975 [234]. Closed genome sequences for a classical *V. cholerae* (O395), and for M66 were subsequently reported [97]. Comparative genomics between these two genome sequences and that of N16961 illustrated the gain and loss of genes as units in these isolates [97], consistent with previous reports [133].

Work by Chun *et al*. was one of the first studies to characterise *V. cholerae* population structures using whole-genome sequencing and comparative genomics [54]. Here, the authors described the phylogenetic relationships between 23 *V. cholerae* isolates, which corroborated the conclusions drawn from molecular studies that M66 was closely-related to the El Tor clone causing the seventh cholera pandemic, and that classical *V. cholerae* were more closely related to the pandemic El Tor clone than were environmental *V. cholerae* [54, 163]. The authors included non-O1 *V. cholerae* in their study, and also included details such as characterising the genes present and absent in each genome, the structures of the CTXφ prophages, and a proposed sequence of genomic island acquisition and loss in the course of the emergence of the clades of El Tor and classical pandemic clades (dubbed pandemic groups, PGs) [54].

Several global phylogenetic studies of cholera epidemiology have since been conducted to describe the routes by which epidemic *V. cholerae* is transmitted globally [158, 189, 234, 235, 308, 309]. Inherent to performing such analyses is the fact that pandemic *V. cholerae* appears to evolve with a very stable molecular clock rate of ≈ 6 substitutions *per* site *per* year [158]. This facilitates the calculation of dated phylogenies with which to infer transmission events

between continents and countries [158, 189, 309]. One of the first such studies was published in 2011, using the genome sequences of 136 *V. cholerae* and dated phylogenies to show that the seventh cholera pandemic had been transmitted globally in three waves [234]. These results built upon molecular and epidemiological data to show that the seventh pandemic of cholera was caused by a single lineage of *V. cholerae* O1 El Tor, and that this lineage appeared to reside in Southeast Asia (*circa* the Bay of Bengal) and to be transmitted in these waves [234]. This *V. cholerae* lineage has since been re-named "7PET", for seventh pandemic El Tor [189].

Other studies, including recent analyses, have explored the contributions of other countries in Asia to these dynamics [308, 310, 311], as well as the micro-evolution of the pandemic lineage in India [200]. Whole-genome sequencing data have been used to provide strong corroborative evidence that the Haitian cholera epidemic had its origin in Nepal [26, 27, 312]. Two studies reported in 2017 demonstrated that since 1970, pandemic cholera in both Africa and Latin America was caused by sub-lineages of 7PET [158, 189]. These data also provided additional phylogenetic support for the transmission of cholera from Nepal or surrounding countries into Haiti in 2020 [189]. Most recently, genomics [309] coupled to epidemiological data [28] have contributed to our understanding of how cholera entered into and spread within Yemen during the ongoing outbreak.

Building on these global analyses, the transmission of 7PET within and amongst households in Dhaka, Bangladesh, a city hyper-endemic for cholera, has been studied. This work found that multiple sub-lineages of 7PET co-existed with one another over the course of four years [235]. This was unlike the dynamics observed in Africa and in global transmission studies [158, 234, 309], which saw sub-lineage replacements rather than the co-circulation of sub-lineages, and it was hypothesised that this reflected the unique setting of a hyper-endemic environment [235]. However, due to insufficient sampling in a single non-endemic country or setting, data were unavailable with which to contrast the observations made in Dhaka, and sub-lineage dynamics in non-hyper-endemic settings remain uncertain. Studying how pandemic *V. cholerae* evolves upon introduction into a naïve setting was a strong focus of the research in this PhD (Aims; section 1.5).

1.4.3 – Patterns of disease and local lineages

For this PhD, one of the most significant observations from the above genomic studies was made as part of the work carried out in Latin America, which led to the identification of lineages of *V. cholerae* O1 El Tor that were distinct from 7PET, but were associated with low levels of disease [189]. These were dubbed 'local lineages'; all were serogroup O1, biotype El Tor, and some (but not all) were of clinical origin and were toxigenic, harbouring CTXφ and *ctxAB*. These local lineages were detectable in Latin America because countries in this region benefitted from surveillance systems which captured cholera epidemics in parallel with sporadic outbreaks of disease caused by *V. cholerae* which were not part of 7PET and did not cause global disease [189]. Outbreaks of disease caused by these local lineages would satisfy the WHO definitions of suspected and confirmed cholera cases (section 1.1.1); however, the patterns of disease caused by these lineages and by bacteria causing sporadic outbreaks of disease are starkly different (Table 1.2):

Pattern of disease	Description	Number of cases	Aetiological agents
Sporadic outbreaks	Can be caused by non-O1/O139 as well as O1 *V. cholerae*	Very few	Very diverse *V. cholerae*, might be either toxigenic or non-toxigenic
Local epidemics	May cause relatively large numbers of cases, but does not transmit as rapidly as pandemic cholera. May be serogroup O1, but not always (*e.g.*, O37, O75, O1 Gulf Coast)	Variable	Local lineages, may be serogroup O1, but not always
Pandemic cholera	Characterised by rapid transmission, large numbers of cases. Cholera, in the epidemiological sense	Hundreds of thousands	7PET and Classical

Table 1.2 – A summary of the three patterns of disease caused by virulent *V. cholerae*. Scheme after [189], and integrating the discussions in this Introduction.

1.5 – Open questions and aims of this thesis

This Introduction has sought to highlight a number of open questions in our understanding of *V. cholerae* biology. Acknowledging that there is more complexity to the *V. cholerae* species

than can be captured by examining pandemic cholera alone, this PhD project was designed to explore the diversity of *V. cholerae* by addressing the following open questions:

a. Although a lot is known about how pandemic 7PET *V. cholerae* evolves on a global scale, we know little about how this pathogen evolves during an epidemic in a single country that is not endemic for cholera.
b. We also have limited information about the genomics of the *V. cholerae* that are present in a country during an outbreak or an epidemic caused by the introduction of 7PET.
c. The relative paucity of non-O1/O139 *V. cholerae* genomes means that we know comparatively little about how virulence genes, drug resistance determinants, plasmids, and other genetic factors of interest are distributed across the *V. cholerae* species more generally than just within 7PET.
d. As local *V. cholerae* O1 lineages begin to be recognised and identified, opportunities begin to present themselves to study why the 7PET (and Classical) lineages cause pandemics, and other lineages do not, even though they might harbour CTXφ and many of the other genetic determinants associated with pandemic cholera.

The aims of this PhD research were framed around these open questions. In order to study these questions, appropriate collections of *V. cholerae* were required. These are described in each of the chapters in this thesis, which has the following four aims:

1. Using nearly 500 strains collected during the 1992-1998 Argentinian cholera epidemic, to characterise how 7PET evolves during an epidemic, and to compare and contrast this to the background of non-epidemic *V. cholerae* isolated from the same places and times as 7PET (Chapter 3).
2. Using long-read sequencing and experimental approaches, to characterise a number of clinically- and historically-important non-O1 *V. cholerae* of clinical origin (Chapter 4).
3. Using the genomes of ~650 diverse *V. cholerae*, to identify the distribution of antimicrobial resistance genes, plasmids, virulence genes, and key genetic determinants of biotype and pathogenicity across diverse *V. cholerae* (Chapter 5).
4. Integrating the knowledge gleaned from Chapters 3-5, to use phylogenetic and genomic information to select rationally a number of live *V. cholerae* for transcriptomic profiling, beginning to examine global gene expression differences amongst pandemic and non-pandemic *V. cholerae* O1 (Chapter 6).