Exploring the genomic and phenotypic diversity of the *Vibrio cholerae* species

Matthew James Dorman

Wellcome Sanger Institute

and

Churchill College, University of Cambridge

July 2020

This dissertation is submitted for the degree of Doctor of Philosophy

Declaration

This dissertation 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.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or 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. I further state that no substantial part of my dissertation 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 Biology Degree Committee (60,000 words).

Matthew J. Dorman
July 2020

Summary

Exploring the genomic and phenotypic diversity of the *Vibrio cholerae* species

Matthew James Dorman

Vibrio cholerae is the aetiological agent of cholera, an acute diarrhoeal disease which is estimated to result in up to 143,000 deaths *per annum*. Cholera is a considerable public health concern because it can spread rapidly in and explosive pandemics. Current pandemic cholera is caused by a highly-clonal phylogenetic lineage of *V. cholerae* serogroup O1, which spreads across the globe in periodic 'waves'. However, *V. cholerae* is a species rich in diversity, and although much is known about the population structure of the pandemic lineages, the biology and pathogenicity of non-pandemic and non-O1 *V. cholerae* has been comparatively neglected. In this dissertation, I have studied the biology, genome dynamics, and diversity of non-pandemic *V. cholerae*, in comparison to the current pandemic lineage.

I first present an analysis of the 1992-1998 cholera epidemic in Argentina, a country which had been free of pandemic cholera for nearly 100 years before 1992. I use the genome sequences of 490 *V. cholerae* from Argentina to study the micro-evolution of the pandemic lineage upon its introduction into a naïve population. I use these data to describe the progression of the Argentinian cholera epidemic using genomic epidemiology approaches, and to contrast this pandemic lineage to the non-epidemic *V. cholerae* that were present in Argentina at the same time as the pandemic lineage.

I then present a study of important recent and historical *V. cholerae* isolates, sequenced to completion using long-read technologies. I describe aspects of these genomes that could only be resolved using closed assemblies, and present functional validations of several *in silico* observations. Having performed this forensic, manual study of a small number of genomes, I then extrapolate those insights into a wider context, by mapping the distribution of key genetic determinants of important *V. cholerae* phenotypes across a phylogenetic tree of 651 highly-diverse *V. cholerae*. Finally, I integrate the knowledge gained in this research to make a rational selection of *V. cholerae* isolates for transcriptomic analysis, based on their phylogenetic position and gene content, to investigate whether differential gene expression might explain the stark differences between pandemic and non-pandemic *V. cholerae*.

The data presented here add substantially to our understanding of the diversity of *V. cholerae*. They emphasise the stark differences in genome flux and evolution between pandemic and non-pandemic lineages. They also show that many of the genetic and phenotypic markers of epidemic and pandemic lineages are misleading, and do not describe that which they were originally chosen to describe.

General contribution, copyright, and ethics statements

The nature of this research project, specifically the collation of a large collection of bacterial strains and DNA extracts, was highly collaborative. Sequencing libraries were generated by the staff of the Wellcome Sanger Institute core pipelines, except where indicated. Electron micrographs were captured by Claire Cormie and David Goulding. In all other cases, I performed the work and analyses described herein, and produced all figures, except where stated in contribution statements at the beginning of each chapter.

Several of the figures and results reported in this dissertation have been published post-peer-review, or are in press, under open-access copyright licences. Any data, figures, or text from manuscripts arising from this research that have been reproduced here are acknowledged and cited in line with the CC-BY 4.0 licence under which they were published. The maps in Chapter 3 rely on data from OpenStreetMap, which are made available under a CC-BY-SA licence which permits reuse with appropriate recognition and citation (see Chapters 2, 3). Figure 3.1 was drawn using publicly-available data from the World Health Organisation and the Pan American Health Organisation. All other figures and data were produced as part of this research and, to our knowledge, should not be affected by matters of external copyright.

All of the bacteria sequenced and handled as part of this PhD were obtained from publicly-accessible culture collections, or were originally collected by our collaborators as part of routine surveillance for public health purposes. Identifiable patient data were not solicited or made available to us at any stage of this research. No experimentation involving eukaryotic cell lines or animal models was performed. Accordingly, ethical approval for these projects was not required.

Publications

Publications arising directly from this work:

Dorman MJ, Domman D, Poklepovich T, Tolley C, Zolezzi G, Kane L, Viñas MR, Panagópulo M, Moroni M, Binsztein N, Caffer MI, Clare S, Dougan G, Salmond GPC, Parkhill J, Campos J, Thomson NR. Genomics of the Argentinian cholera epidemic elucidate the contrasting dynamics of epidemic and endemic *Vibrio cholerae*. *Nature Communications* **11** (1):4918. PMCID: PMC7530988. DOI: 10.1038/s41467-020-18647-7.

Dorman MJ, Kane L, Domman D, Turnbull JD, Cormie C, Fazal M-A, Goulding DA, Russell JE, Alexander S & Thomson NR (2019). The history, genome and biology of NCTC 30: a non-pandemic *Vibrio cholerae* isolate from World War One. *Proceedings of the Royal Society B* **286** (1900): 20182025. PMCID: PMC6501683. DOI: 10.1098/rspb.2018.2025.

Dorman MJ*, Domman D*, Uddin MI*, Sharmin S, Afrad MH, Begum YA, Qadri F & Thomson NR (2019). High quality reference genomes for toxigenic and non-toxigenic *Vibrio cholerae* serogroup O139. *Scientific Reports* **9** (1): 5865. PMCID: PMC6458141. DOI: 10.1038/s41598-019-41883-x. (* Joint first author)

Dorman MJ & Thomson NR (2020). Community evolution: Laboratory strains in the era of genomics. *Microbiology* **166** (3): 233-238. Invited "insight review" article. PMCID: PMC7376263. DOI: 10.1099/mic.0.000869.

All remaining data from this dissertation are in preparation for peer-reviewed publication.

Acknowledgements

I thank Nick Thomson very sincerely for supervising this PhD research, and for his advice, guidance, and patience throughout this project. I also thank my thesis committee members – George Salmond and Gordon Dougan, my co-supervisors, and Julian Parkhill, my graduate advisor – for their input and suggestions throughout this research.

I acknowledge formally the support of those who have contributed materials, strains, and genome sequences to this project, particularly Josefina Campos and the team at the Malbrán Institute, Sarah Alexander and the team at NCTC, Florian Marks (International Vaccine Institute), Claire Jenkins (Public Health England), and Firdausi Qadri (icddr,b). I am very grateful to have had the opportunity to train in Argentina with the *Enterobacterias* team at the Malbrán Institute, and to have been able to visit NCTC to collate metadata and observe the NCTC operation. I thank all of those named in each chapter's contribution section for their support, and I particularly thank Daryl Domman for his intellectual input into the design of this work, and both Leanne Kane and Charlotte Tolley for their help working at CL3.

The composition of Nick's research group has changed considerably and frequently over the last four years. Rather than risking forgetting to thank anybody by name and causing inadvertent offence, I will instead thank everybody that has been part of the group or in our offices since 2016, and at the Sanger Institute more broadly, who has been involved in this work and has made suggestions or comments throughout the process. I also thank those at Sanger who played an equally important role by carrying out DNA sequencing, and providing the administrative support, infrastructure, facilities and services necessary to see this work completed, particularly Sally Kay and Liz McMinn, all of the Pathogen Informatics group, our Programme's admin team including Danielle Walker, Joseph Woolfolk, and Kate Auger, and the Institute's Graduate Studies office.

Lastly, I acknowledge Wellcome, for financial support.

Table of contents

List of figures

List of tables

Abbreviations

7PET	Seventh pandemic El Tor
AMR	Antimicrobial resistance
ANI	Average nucleotide identity
ANLIS	Administración Nacional de Laboratorios e Institutos de Salud
AST	Antimicrobial Sensitivity Test
ATCC	American Type Culture Collection
ATCSA	Anti-Terrorism, Crime and Security Act
ATP	Adenosine triphosphate
AWD	Acute watery diarrhoea
bp	Base pair
cAMP	Cyclic adenosine monophosphate
CDC	Centers for Disease Control and Prevention
CDS	Coding sequence
CFTR	Cystic fibrosis transmembrane conductance regulator
Chr1	Chromosome One
Chr2	Chromosome Two
CL3	Containment Level Three
Conc.	Concentration
CT	Cholera toxin
CTXφ	CTX bacteriophage
ddH_2O	Double distilled water
DNA	Deoxyribonucleic acid
E. coli	*Escherichia coli*
EDTA	Ethylenediaminetetraacetic acid
ENA	European Nucleotide Archive
ESBL	Extended-spectrum β-lactamase
FDR	False discovery rate
GAVI	Global Alliance for Vaccines and Immunisation
gDNA	Genomic deoxyribonucleic acid
GPS	Global Positioning System
Gsp	General secretory pathway
GTFCC	Global Taskforce for Cholera Control

GTP	Guanosine triphosphate
HGT	Horizontal gene transfer
icddr,b	International Centre for Diarrhoeal Disease Research, Bangladesh
Inc	Incompatibility group
INEI	Instituto Nacional de Enfermedades Infecciosas
INIDEP	Instituto Nacional de Investigación y Desarrollo Pesquero
IVI	International Vaccine Institute
LAT	Latin American Transfer
LPS	Lipopolysaccharide
MARTX	Multifunctional autoprocessing RTX
MIC	Minimum Inhibitory Concentration
MSC	Microbiological Safety Cabinet
MSF	Médecins Sans Frontières
MSHA	Mannose-sensitive haemagglutinin
NEB	New England Biolabs
NCTC	National Collection of Type Cultures
OCV	Oral cholera vaccine
ORF	Open reading frame
ORS	Oral Rehydration Solution
PAHO	Pan American Health Organisation
PCA	Principal component analysis
PCR	Polymerase chain reaction
PE	Paired-end
PFGE	Pulse-field gel electrophoresis
PG	Pandemic Group
PHE	Public Health England
ppGpp	Guanosine tetraphosphate
(p)ppGpp	Guanosine pentaphosphate
RNA	Ribonucleic acid
RNA-seq	Ribonucleic acid sequencing
rRNA	Ribosomal ribonucleic acid
RTX	Repeats-in-toxin
SEM	Scanning electron micrograph
SXT (SXT/R391)	Sulfamethoxazole and trimethoprim resistant conjugative element

T1SS	Type I secretion system
T2SS	Type II secretion system
T3SS	Type III secretion system
T6SS	Type VI secretion system
TCBS	Thiosulfate-citrate bile salt
TCP	Toxin co-regulated pilus
TEM	Transmission electron microscopy
TIGR	The Institute for Genomic Research
USA	United States of America
V. cholerae	*Vibrio cholerae*
V. metoecus	*Vibrio metoecus*
WHO	World Health Organisation
WSI	Wellcome Sanger Institute
WW1	World War One
VPI-1	*Vibrio* pathogenicity island 1
VPI-2	*Vibrio* pathogenicity island 2
VSP-1	*Vibrio* seventh pandemic island 1
VSP-1	*Vibrio* seventh pandemic island 1

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].

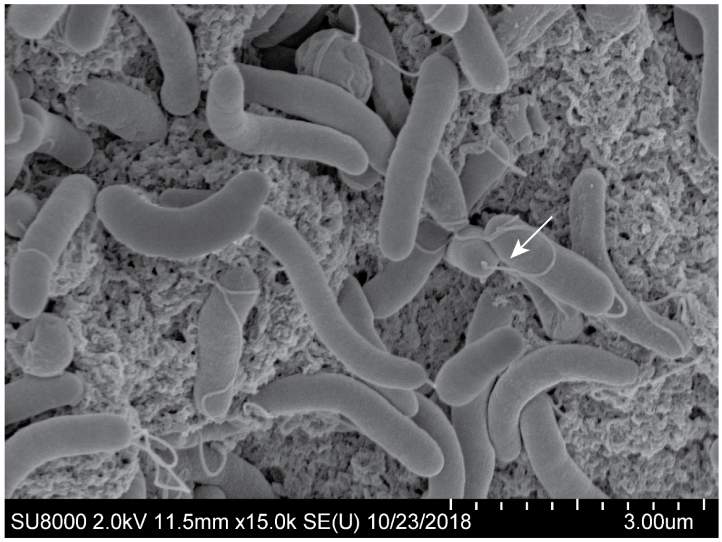


Figure 1.1 – Scanning electron micrograph of *V. cholerae*. SEM produced from fixed *V. cholerae* colonies grown on LB agar plates. Several comma-shaped bacteria can be seen in this image, amongst extracellular polysaccharide (biofilm). A cell producing a polar monotrichous flagellum is indicated (white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

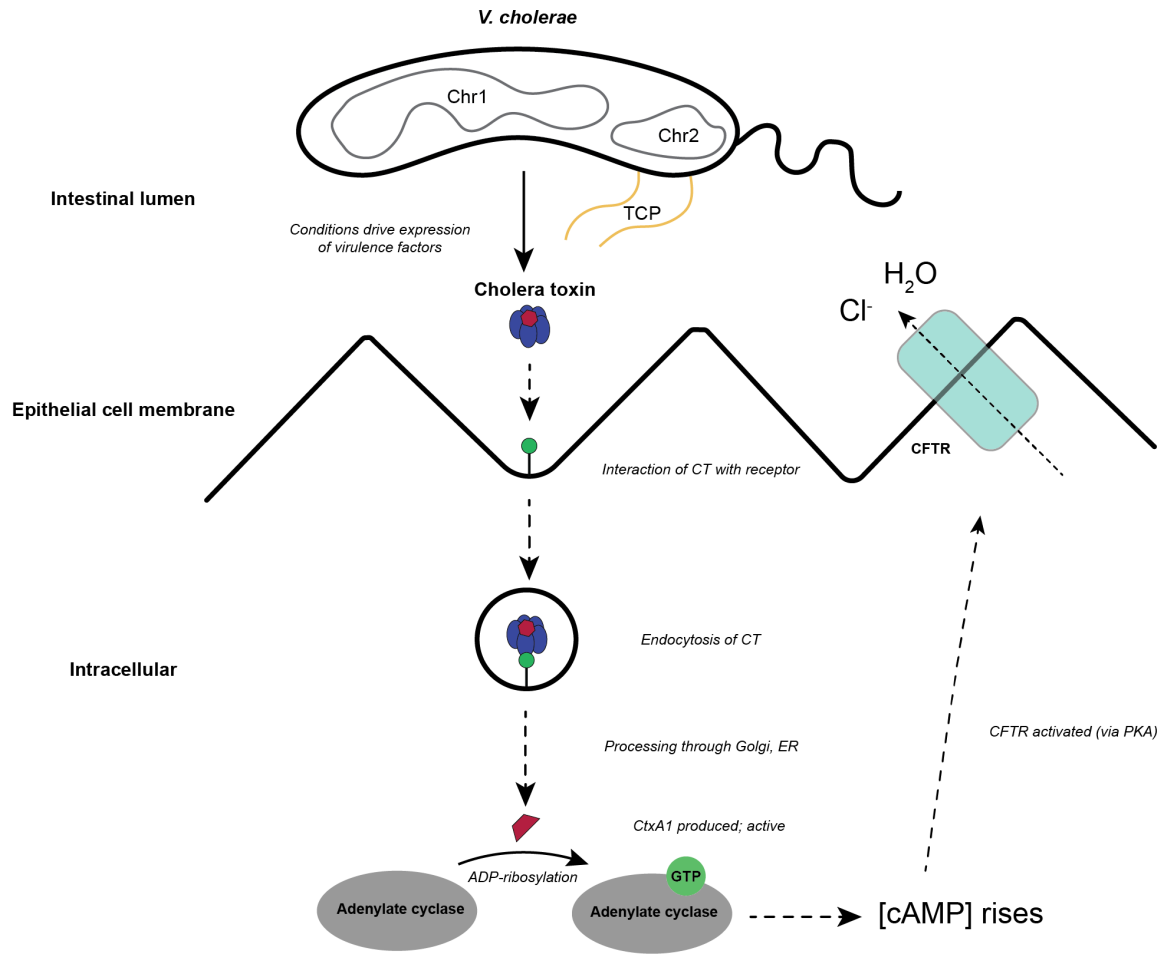


Figure 1.2 – Model of *V. cholerae* pathogenesis leading to the diarrhoea characteristic of cholera. Drawn from steps outlined in section 1.2.2. CT receptors (G_{M1} or others) are indicated as a green circle on the epithelial membrane. TCP is elaborated by *V. cholerae* and acts as a colonisation factor 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).

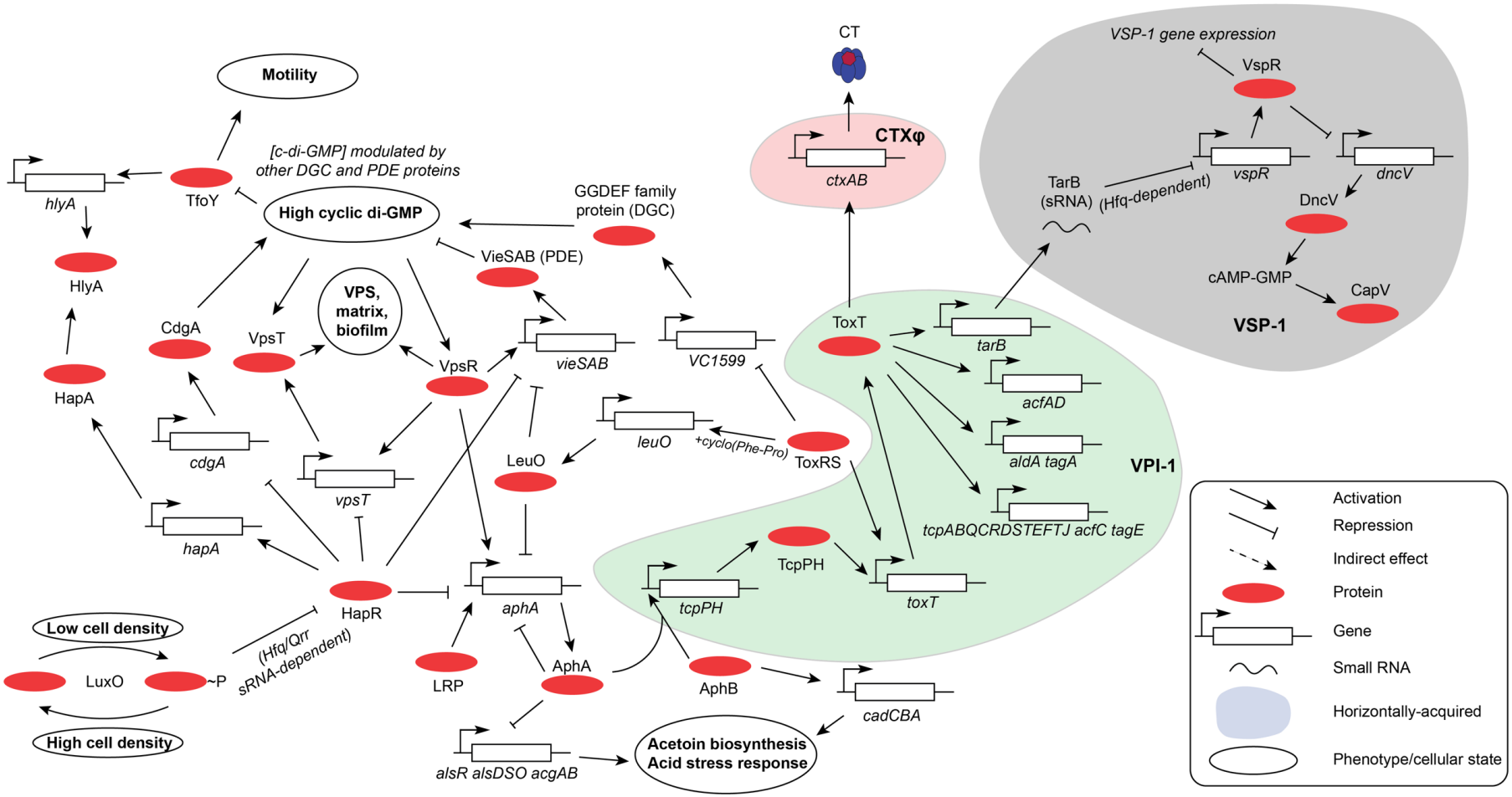


Figure 1.3 – Summary of gene/protein interactions involved in regulating virulence gene expression. This model describes those genes mentioned in this thesis and is not exhaustive. The input of quorum-sensing into this system has been simplified for presentation purposes. Genes that are encoded on mobile 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).

Chapter 2

Methods

2.1 – Computational analyses

2.1.1 – Bacteria sequenced in the course of this PhD research

A list of *V. cholerae* that were sequenced in the course of this study is provided in Appendix 1. Additional previously-published genomes were also collated for analyses in Chapters 3-5. If the short-reads for externally-sequenced genomes were available, these were included with the short-reads generated in this study for analysis. If these external genomes were only available as assemblies and short-reads were required (*e.g.*, for mapping analyses), paired-end 100 bp Illumina reads were simulated from assemblies using wgsim v.0.3.2 [1], part of SAMtools [313], and parameters '-e 0 -r 0 -X 0 -1 100 -2 100' (these settings assume a haploid genome and are appropriate for bacteria). Additional *V. cholerae* genomes were also downloaded from Refseq as assemblies.

2.1.2 – Genome assembly

2.1.2.1 – Short-read data

Throughout this PhD research, short-read sequencing data generated at WSI were assembled using SPAdes v3.8.2 [314] as part of a WSI high-throughput analysis pipeline [315]. Externally-generated sequencing reads were assembled using the same pipeline.

2.1.2.2 – Long read data

The *V. cholerae* O139 long-read assemblies described in Chapter 4 (section 4.3.1) were produced using HGAP v3 and SMRT analysis software v2.3.0 [316]. The fold coverage to target when picking the minimum fragment length for assembly was set to 30 and the approximate genome size was set to 3 Mbp. Assemblies were circularised using Circlator v1.1.3 [317] and the pre-assembled reads (also known as corrected reads). Circularised assemblies were polished using the PacBio RS_Resequencing protocol and Quiver v1 [316]. Pilon v1.19 [318] did not identify any single nucleotide variants (SNVs) in any of the PacBio

[1] https://github.com/lh3/wgsim

assemblies using the corresponding short-read data – accordingly, no short-read corrections were made to these assemblies.

The NCTC 30 genome (Chapter 4, section 4.3.7) was similarly assembled from PacBio RSII reads using HGAP v3 and the RS_HGAP_Assembly.2 protocol, performed using SMRT Portal and SMRT Analysis v2.3.0.140936.p5.167094 [316]. Minimum polymerase read quality and length were set to 0.8 and 100, respectively. A minimum length of 500 bases was used to filter sub-reads. The minimum seed read length for assembly was set to 6,000, with BLASR options '-noSplitSubreads -minReadLength 200 -maxScore -1000 -maxLCPLength 16', an expected genome size of 5 Mbp, and a target coverage of 30 X. Contigs were circularised using Circlator v1.5.3 [317] using both the assembly and the corrected reads, and a final assembly was obtained by using the circularised sequences as a reference to re-assemble the PacBio reads using the RS_Resequencing.1 protocol (minimum subread length = 50 bp, minimum polymerase read quality = 75%, minimum polymerase read length = 50 bp, BLASR maximum divergence = 30%, minimum anchor size = 12). This assembly was then corrected using Quiver v1. The finished assembly was covered to an average depth of 148.01 X. To check the accuracy of the PacBio assembly, the corresponding Illumina short-reads were mapped to the assembly using SMALT v0.5.8 [2] (maximum insert size = 1000, minimum insert size = 50). No SNVs were identified in the assembly upon mapping of these data.

2.1.2.3 – Hybrid assemblies and data visualisation

The hybrid assemblies described in Chapter 5 (section 5.3.7) were produced using both the long- and short-reads for each isolate and Unicycler v0.4.0 [319], with options '-t 4 --mode conservative'. The De Brujin graphs from these assemblies were visualised using Bandage v0.8.0 [320].

2.1.3 – Sequencing quality control

Contamination of genome sequences was assessed using several methods. Kraken [321] was used to assess whether sequences belonging to bacterial species other than the species of interest were present in the sequencing data for an isolate. Additionally, the length of genome

[2] http://www.sanger.ac.uk/science/tools/smalt-0

assemblies was inspected and summarised using assembly-stats v1.0.1 [3]. Assemblies that were substantially larger than expected for a *V. cholerae* genome were treated as potentially-contaminated (*i.e.*, assembly length was ≥ 5 Mbp; the *V. cholerae* reference genome is 4.1 Mbp in length [59]). Phylogenetic trees were also inspected at the beginning of each analysis, to identify stark outliers (*e.g.*, isolates on very long branches). Any sequences which were suspected to be of poor quality were excluded from downstream analyses.

2.1.4 – Automated genome annotation

All genome assemblies used in this research were uniformly annotated using Prokka v1.5 [322] and a genus-specific database for *Vibrio* [323]. This was either run as part of the WSI automated assembly pipeline (for all automatically-assembled genomes, for genomes downloaded from RefSeq, and for any assemblies that were produced from simulated reads), or manually (for hybrid and long-read-only assemblies generated in Chapters 4 and 5).

2.1.5 – Pangenome calculations

All pangenomes described in this thesis were generated using Roary v3.12.0 [324], run with options '-e --mafft -s -cd 97'. These settings prevent the splitting of paralogous genes into separate gene clusters to avoid the over-partitioning of orthologous genes into separate clusters, and impose a CD-HIT clustering cut-off of 97%. Prokka-annotated genome assemblies in GFF3 format were used as input for these calculations. Gene presence/absence matrices were visualised using roary_plots.py v0.1.0 [4].

2.1.6 – Short-read mapping and SNV detection

High-quality SNVs, including small indels, were identified as described by Harris *et al* [325] using SMALT v0.7.4 [5]. Briefly, variant sites was identified from mapped sequence data using SAMtools mpileup v0.1.19 [313] and parameters '-d 1000 -DSugBf', and bcftools v0.1.19 [313], to produce a BCF file of all variant sites from which high-quality SNVs were identified using previously-described quality thresholds [325]. For analyses of 7PET genomes, mapping

[3] https://github.com/sanger-pathogens/assembly-stats
[4] https://github.com/sanger-pathogens/Roary/tree/master/contrib/roary_plots
[5] http://www.sanger.ac.uk/science/tools/smalt-0

and SNV-calling were performed against the N16961 reference genome sequence (accession # LT907989/LT907990 [59]). For the analysis of LAT-1 described in Chapter 3 (section 3.4.5), mapping and SNV-calling were performed against the sequence of the LAT-1 isolate A1552 (accession # CP025936/CP025937 [326]).

2.1.7 – Detection of potentially-recombined chromosomal regions

Potentially-recombined regions of *V. cholerae* genomes were identified using Gubbins v1.4.10 [327] and a "pseudo-genome" multiple-sequence alignment of all variant and invariant sites relative to a reference genome sequence. This tool produced a variant-only alignment of SNVs which excluded all variants predicted to be located within recombined regions of the chromosome. This alignment was used for subsequent phylogenetic analyses, to avoid SNVs in recombined regions potentially interfering with the topology of phylogenetic trees.

2.1.8 – Phylogenetic analyses and tree-building

7PET and LAT-1 phylogenies were calculated from non-recombinant SNVs called against the N16961 and A1552 reference sequences by mapping, as described above (section 2.1.6 and 2.1.7). *V. cholerae* species phylogenies were calculated from the core-gene alignment generated by Roary (section 2.1.5); core-gene alignments were trimmed using trimAl v1.4.1 [328], and SNP-sites v2.4.1 and v2.5.1 [329] was used to extract the variant sites from this trimmed alignment. The sizes of alignments used for calculations are reported throughout the thesis.

The maximum-likelihood phylogenetic trees presented in this thesis were calculated using IQ-Tree v1.6.10 [330] unless otherwise stated in the text, and either the alignments of non-recombinant SNVs produced by Gubbins (section 2.1.7) or the SNV-only alignments from a core-gene alignment, described above. Trees were calculated under the general time reversible (GTR) and ascertainment bias correction (ASC) models [331]. In order to assess the robustness of the computed phylogenies, five thousand approximate likelihood ratio tests [332] and ultrafast bootstrap approximations [333] were performed for each tree calculation.

2.1.9 – Data clustering and lineage assignment

In order to identify clusters of similar sequences from multiple-sequence alignments, alignments of parsimony-informative SNVs were extracted from SNV-only alignments using extract_PI_SNPs.py [6] (*i.e.*, any SNVs that were private to a single isolate were removed from the alignment). For alignments that were derived from *V. cholerae* species core gene alignments, sequence data for the *Vibrio* spp. outgroup were deleted from the alignment prior to running extract_PI_SNPs.py.

Alignments of parsimony-informative SNVs were used for cluster analysis using the R implementation of Fastbaps v1.0.1 [334]. The Dirichlet Process Mixture model was used to cluster 7PET and LAT-1 genomes, whereas the Bayesian Hierarchical Clustering prior [335] was used to cluster sequences in *V. cholerae* species phylogenies.

2.1.10 – Comparative genomics

Synteny plots were generated using ACT v13 [336] and Easyfig v2.2.2 [337] to enable comparative genomics analyses. Both of these tools rely on BLASTn [338] comparisons between two genome assemblies. Cut-offs for Easyfig comparisons were set to a maximum *e*-value of 0.001 and minimum length of 0; the minimum BLASTn identity percentage is specified in each comparison figure throughout the thesis. BamView [339] was used to visualise the mapping of Illumina reads to a genome sequence using the BAM file of mapped sequencing data.

2.1.11 – Pairwise SNV and ANI calculation

SNV distance matrices were calculated from SNV-only alignments using snp-dists v0.4 [7]. Average nucleotide identity (ANI) values were calculated using FastANI v1.0 [340] and a reference sequence as stated in the text.

[6] https://gist.github.com/jasonsahl/9306cd014b63cae12154
[7] https://github.com/tseemann/snp-dists

2.1.12 – Assessment of temporal signal in phylogenetic data

Maximum-likelihood phylogenetic trees produced in section 2.1.8 were read into TempEST v1.5.1 [341] to assess whether a temporal signal could be detected within the phylogeny. The names of leaves on the phylogeny were amended to include the date of isolation, where known.

2.1.13 – Detection of antimicrobial resistance genes, wbeT *and* ctxB *variants, plasmid replicons and virulence determinants*

Antimicrobial resistance genes and plasmid replicons were detected in genomes using both ARIBA v2.12.1 [342] and ABRicate v1.0.1 [8]. ARIBA performs local assembly of sequence data against a reference database of genes of interest, and was therefore only used when short-reads were available for all genomes of interest. In contrast, ABRicate screens assembled contigs for genes of interest, and was used when assemblies were available for all genomes of interest. The ResFinder database of antimicrobial resistance genes [343], and the PlasmidFinder database of plasmid replicon sequences [344] were used for ARIBA and ABRicate analyses (both databases accessed on 23/06/2019 and 11/05/2020 for ARIBA and ABRicate analyses, respectively). The VFDB database (accessed on 11/05/2020) was used for identifying virulence genes of interest using ABRicate. Default cut-offs for identity percentage and length were used throughout this thesis unless otherwise specified in the text.

ARIBA was also used to assign *wbeT* and *ctxB* genotypes to *V. cholerae* isolates. This relied on a custom database consisting of the *ctxB* nucleotide sequence from N16961 (accession # LT907989/LT907990) and the intact *wbeT* sequence from NCTC 9420 [213] (accession # CP013319/CP013320). This *wbeT* sequence translates into a protein sequence which is 100% identical to the WbeT sequence from the Ogawa isolate VX44945 (AEN80191.1 [196]), which has been used for *wbeT* genotyping in previous studies [158, 189].

The NCTC 30 genome assembly was screened for antimicrobial resistance genes using the ResFinder web server [343] (database version 2018-02-19) using default cut-offs of 90% nucleotide identity and 60% minimum length.

[8] https://github.com/tseemann/abricate

2.1.14 – Confirming the presence of genomic elements

The presence of the WASA-1 genomic island, a marker characteristic of the LAT-1 sub-lineage [234, 235], was confirmed in genome assemblies by using BLASTn to query the assembly for the presence of the WASA-1 sequence, as well as by using pangenome gene presence/absence matrices to check for the presence of WASA-1 gene orthologues in an assembly of interest. Similarly, the presence and absence of the *V. cholerae* serogroup O1 biosynthesis operon was confirmed using the pangenome gene presence/absence matrix, and by testing for the presence of the O1 biosynthesis operon sequence using BLASTn.

2.1.15 – Construction of BLAST atlases

BLAST atlas figures were generated using annotated genome assemblies (EMBL format) and the GView web server [9], which relies on CGView [345]. The GC% and GC skew calculation for each comparison was also carried out using GView.

2.1.16 – Statistical tests

Statistical tests and p-value cut-offs are listed and described in the text of the thesis. Unless otherwise specified, statistical tests were performed using the R package, v3.5.1 [346]

2.1.17 – Promoter sequence analysis

The *tcpPH* and *aphA* promoter sequences were extracted from genome assemblies using *in silico* PCR, a custom script at WSI which searches for the sequences of 5' and 3' "primers" within an assembly and reports the sequence between those as output. To "amplify" P_{tcpPH}, the primer sequences 5'-`cgtaaggggcaaaatgtcacaggaa`-3' and 5'-`atagacttgattagtgcattcattc`-3' were used (maximum amplicon length 1,500 bp, minimum length 100 bp), and for P_{aphA}, 5'-`gaatgcgcaatactggttaa`-3' and 5'-`catcgacaggtttggcttgg`-3' (maximum length 2,000 bp; minimum length 100 bp). Up to three mismatches were permitted between primer and assembly sequences. The extracted promoter sequences were concatenated into a multi-FASTA file and aligned using Clustal

[9] https://server.gview.ca/

Omega [347] and Seaview v4.6.1 [348]. trimAl v1.4.1 [328] was used to extract 50 bp sequences from this alignment which were then used to identify motifs and produce nucleotide usage frequency figures using MEME v5.1.1 [349]

2.1.18 – RNA-seq data analysis and statistical cut-offs

The Rockhopper 2 package [350, 351] was used to calculate expression of genes using the sequence and annotation of the N16961 isolate as a reference. Genes were determined to be differentially expressed if the $\log_2$(fold change) of a gene's expression in one condition relative to a second condition was greater than or equal to 2, or less than or equal to -2. Additionally, the false discovery rate (FDR) q-value cut-off for each result was set to <0.01 – this was used rather than a p-value cut-off in order to account for multiple testing implicit in large RNA-seq comparisons. Principal component analyses (PCA) were calculated using prcomp in R v3.5.1 and a matrix of normalised expression values *per* gene for all replicates in an experiment.

2.1.19 – Additional data visualisation

Incidence data reported in Chapter 3 (*e.g.*, section 3.4.2) were visualised and annotated using Tableau Desktop 2018.31. Maps produced in Tableau made use of OpenStreetMap (© OpenStreetMap contributors) which is licenced under a CC-BY-SA licence [10]. Phylogenetic trees were visualised and annotated using Figtree v1.4.3 [11], iCANDY [12], iTOL v3 [352], and the Phandango web server [353]. Various graphs and plots were produced using Adobe Illustrator CC v23.0.4, or R v3.5.1 with the ggplot2 v3.1.1 package [354] and reshape v0.8.8 [355] packages. Various comparative genomics figures were generated using Artemis v16 [356], ACT v13 [336], DNAplotter v1.11 [357], and Easyfig v2.2.2 [337]. Where figures were edited manually, this was performed using Adobe Illustrator CC v23.0.4.

[10] https://www.openstreetmap.org/copyright
[11] http://tree.bio.ed.ac.uk/software/figtree/
[12] https://github.com/simonrharris/iCANDY

2.2 – Experimental methods

2.2.1 – Strains, plasmids and oligonucleotides

The bacterial strains, plasmids, and oligonucleotide primers used in thesis chapters involving molecular and genetic experiments (Chapters 4 and 6) are listed in Table 2.1. This list excludes isolates which were exclusively used for whole-genome sequencing – these are listed in Appendix 1.

Internal ID	Strain name	Genotype/Details	Source/Reference
Vibrio cholerae			
MJD382	NCTC 30	Non-O1/O139 (probably O2). Isolated in 1916; Alexandria, Egypt. AmpR	NCTC
MJD439		Second clone of NCTC 30. AmpR	
MJD367	NCTC 10732	Serotype O1 Inaba, classical biotype. Isolated in 1952; India.	NCTC
MJD1404			
MJD389	NCTC 5395	Serotype O1 Ogawa, El Tor biotype. Pre-seventh pandemic. Isolated in 1938; Iraq. AmpS	NCTC. Sequenced in 2016 [213]
MJD1402	NCTC 10256	O1 El Tor biotype. 7PET lineage. Toxigenic. Hong Kong, 1961.	NCTC [358]
MJD1403	MJD474	O1 El Tor biotype. 7PET lineage. Toxigenic. Traveller from Somalia, 2017.	Claire Jenkins
MJD1405	NCTC 5596	Classical lineage. O1 classical biotype. Pre-1939.	NCTC
MJD1406	A213	Gulf Coast. O1 El Tor biotype. Non-toxigenic. Georgia, 1984.	IVI [234]
MJD1407	A219	Gulf Coast. O1 El Tor biotype. Toxigenic. Georgia, 1986.	IVI [234]
MJD1408	MJD462	Non-O1. Related to MS-6 and Chinese non-7PET isolates. Traveller from Thailand, 2017.	Claire Jenkins
MJD1409	NCTC 9422	O1 El Tor biotype. Non-7PET. Non-toxigenic. Pre-1955.	NCTC
Escherichia coli			
MJD839	ER2420 pACYC184	K-12 cloning strain harbouring pACYC184. CmR TcR	Francesca Short/NEB
MJD841	NEB® 5-alpha	*fhuA2 Δ(argF-lacZ)U169 phoA glnV44* Φ80 *Δ(lacZ)M15 gyrA96 recA1 relA1 endA1 thi-1 hsdR17*	NEB
MJD842	NEB® 5-alpha pUC19	K-12 cloning strain harbouring pUC19. AmpR	This study
MJD844	NEB® 5-alpha pACYC184	K-12 cloning strain harbouring pACYC184. CmR TcR	This study
MJD847	MJD847	NEB® 5-alpha harbouring pMJD61. AmpR CmR TcS	This study

Plasmid name	Genotype/Details	Source/Reference
pACYC184	Low-copy cloning vector. CmR TcR	[359]
pUC19	High-copy cloning vector, ampicillin resistance positive control. AmpR	[360]
pMJD61	pACYC184 Ω(*tet*::$bla_{CARB\text{-}like}$). AmpR CmR TcS	This study

Primer ID	Other name	Sequence 5'-3'
oMJD96	BamHI_blaCARB-like-NCTC30_orf_5	CCGGATCCGGTTTCAGTGCCTAATGCTTTAAGTTAAGATG
oMJD97	blaCARB-like-NCTC30_orf_SalI_3	CCGTCGACATCAACGCGACTGTGATGTATAAACTTCAA
oMJD88	blaCARB-like-NCTC30_int_5	TGGGGTCACATACATGAAGTCT
oMJD89	blaCARB-like-NCTC30_int_3	CAGCAATACTCCACTTCACTG
oMJD98	pACYC184_tet_seq_Pf	GTTAAATTGCTAACGCAGTC
oMJD99	pACYC184_tet_seq_Pr	GTGAATCCGTTAGCGAGGTG
oMJD135	VC_2135_check_Pf	GTCAGGCAGATAGCTCAAACT
oMJD136	VC_2135_check_Pr	CTCATTGCTACCTCTGATGCC

Table 2.1 – Strains, plasmids, and oligonucleotides. Restriction enzyme recognition sites are underlined. AmpR: ampicillin resistant; CmR: chloramphenicol resistant; TcR: tetracycline resistant. AmpS: ampicillin sensitive. TcS: tetracycline sensitive.

2.2.2 – Routine bacterial culturing

V. cholerae and *E. coli* were cultured routinely on LB media (10 g/L tryptone, 5 g/L yeast extract, 5 g/L sodium chloride), supplemented with agar (1.5%) and antibiotics as appropriate (100 µg/ml ampicillin, 10 µg/ml chloramphenicol, or 10 µg/ml tetracycline). Pre-poured thiosulfate-citrate bile salt (TCBS) agar plates (Oxoid # PO0194) was used for selective culture of *V. cholerae* (on this medium, *V. cholerae* typically grows as yellow colonies). Plasmids were maintained in strains by culturing on LB media supplemented with antibiotics where required. All manipulations with live *V. cholerae* were carried out at Containment Level 3 (CL3) within a Class II microbiological safety cabinet (MSC).

2.2.3 – Single-colony purification of strains received during this project

Throughout this project, live *V. cholerae* were received from multiple sources, typically shipped on agar stabs or slopes. Upon receipt, these isolates were sub-cultured immediately on to LB agar plates and grown overnight at 37 ˚C (passage 1). At least two distinctly-isolated and representative colonies were streaked onto both TCBS and LB agar plates and grown overnight at 37 ˚C (passage 2). If bacteria recovered well on TCBS media, single colonies were taken from this plate and purified once more onto LB media (grown overnight at 37 ˚C). Colonies from these plates (passage 3) were then taken and mixed in a cryotube containing a 1:1 mix of 50% v/v glycerol and sterile LB media. These were indexed and frozen at -80 ˚C. If growth on TCBS was poor on day 1, distinctly-isolated colonies were taken from the LB agar plates from day 1 and frozen similarly. At least two independent colonies were frozen per strain.

2.2.4 – Recovery of lyophilised NCTC isolates

V. cholerae were received from NCTC as lyophilised stocks, and were revived, sub-cultured, and frozen at -80 ˚C following the method published by Public Health England Culture Collections [13]. Ampoules containing lyophilised bacterial culture were opened in a Class II MSC, and the contents were rehydrated at room temperature with 0.5 ml LB broth. After five minutes incubation, this suspension was mixed and spread over LB agar plates, which were incubated overnight at 37 ˚C (passage 1). Three well-isolated colonies from these plates were

[13] https://www.phe-culturecollections.org.uk/

single-colony purified onto both LB and TCBS agar (passage 2). Colonies were taken from TCBS plates (or from LB plates if the strain grew poorly on TCBS plates) and were used to inoculate 3 ml LB liquid media in a disposable test-tube. This was incubated for 24 h with shaking (180 rpm) at 37 ˚C (passage 3). Cultures were mixed with glycerol (25% v/v final concentration) and frozen at -80 ˚C.

2.2.5 – Extraction of gDNA from V. cholerae

DNA extractions from *V. cholerae* were carried out using the Masterpure Complete DNA and RNA Purification kit (Epicentre, #MC85200), with modifications to the manufacturer's protocol. An LB agar plate was lawned using a single isolated colony from each strain to be processed. This plate was grown overnight (37 ˚C). Five loopfuls of bacterial culture were taken from this plate and mixed with 300 µl Tissue & Cell Lysis Solution containing Proteinase K, vortexed (10 s), and then incubated at 65 ˚C for the shorter of 20-25 min or until the suspension had fully cleared. Samples were vortexed every ~ 5 min during this incubation to ensure even lysis of bacterial cells. RNAse A was then added to each sample, which were incubated at 37 ˚C for 30 min to ensure complete digestion of residual RNA before samples were chilled on ice for ≥ 5 min.

Proteins were precipitated from each sample by adding MPC Protein Precipitation Solution (150 µl) and vortexing the sample immediately (10 s). Precipitated debris was collected by centrifugation (16,000 x *g*; 10 min; 4 ˚C). Additional MPC reagent was added to each cleared sample (30 µl) to remove residual protein contaminants. Samples were then vortexed and re-centrifuged (16,000 x *g*; 10 min; 4 ˚C). Genomic DNA (gDNA) was precipitated from the cleared supernatant in a fresh 1.5 ml Eppendorf tube using room-temperature isopropanol (500 µl). Precipitated gDNA was collected by centrifugation (16,000 x *g*; 10 min; 4 ˚C) and washed with room-temperature 70% v/v ethanol (2 x 1 ml). The pellet was dried and resuspended in 80 µl nuclease-free water – ethylenediaminetetraacetic acid (EDTA) was excluded from the resuspension solution to avoid interfering with PacBio sequencing chemistry. Samples were then tested for sterility: 2 µl was spotted onto an LB agar plate which was incubated overnight at 37 ˚C; the absence of bacterial growth was interpreted as the sample being sterile. Sterile DNA samples were removed from containment and, where appropriate, submitted for sequencing.

2.2.6 – Next-generation DNA sequencing

gDNA was sequenced using the Illumina X10 and the PacBio RSII platforms by the high-throughput and long-read teams at WSI. Briefly, DNA fragments of approximately 450 bp were produced from 0.5 µg gDNA and used for Illumina library creation. Libraries were sequenced on a 150 bp paired-end run (target coverage 30 X; *V. cholerae* were multiplexed with up to 384 samples per Illumina X10 lane). Approximately 10 µg gDNA was used for PacBio sequencing, using polymerase version P6 and C4 sequencing chemistry reagents. gDNA from NCTC 30 batch 4 was sequenced on the PacBio Sequel platform.

2.2.7 – High-throughput analysis of growth kinetics

Single colonies of *V. cholerae* were picked and suspended in 0.5 ml LB broth by vortexing (10 s). Two microlitres of this suspension were used to inoculate 150 µl LB in a 96-well flat-bottomed microtitre plate (Corning CoStar #3595) and a gas-permeable seal was applied to the plate. Plates were incubated at 37 ˚C in a BMG Fluostar Omega microtitre plate reader for 24 h, with shaking at 300 rpm using the 'meander corner well shaking' mode. Absorbance (λ = 600 nm) was measured every 15 min (cycle time: 900 s); four readings were taken *per* well *per* timepoint in a square matrix and averaged. This was to minimise the risk that debris or aggregated cells passing through the light beam would artificially increase absorbance readings. Three colonies *per* strain (biological triplicate) were used to inoculate six technical replicate cultures each.

2.2.8 – Transmission electron microscopy

Bacterial colonies were picked and suspended in 0.5 ml sterile water. The suspension (4 µl) was applied to a glow-discharged Formvar carbon film copper TEM grid (FCF-100-Cu) and mixed with ammonium molybdate solution (2.5% final concentration). Following initial tests to confirm that this treatment killed *V. cholerae*, inoculated grids were removed from CL3 for microscopic analysis. Images were acquired using a FEI Tecnai G2 Spirit BioTWIN.

2.2.9 – Motility assays

Clearly-isolated bacterial colonies were re-suspended in 0.5 ml LB media by vortexing in a Class II MSC. The suspension (2 µl) was used to inoculate motility LB agar plates (0.3% agar in 140 mm dishes). The pipette tip was pushed through the agar surface during inoculation. Plates were incubated face up at 37 ˚C for 18 h prior to being photographed. Each experiment was repeated in biological triplicate.

2.2.10 – Amplicon sequencing of the flrC *locus*

The *flrC* mutation was confirmed by amplifying *flrC* from *V. cholerae* gDNA using Phusion® high-fidelity DNA polymerase (NEB, # M0530S) and primers oMJD135 and oMJD136 (annealing temperature: 55 ˚C; extension time: 60 s, 30 cycles). The resultant amplicon was purified using the QIAquick PCR Purification kit (Qiagen, #28104) according to the manufacturer's instructions. The DNA concentration in each purified sample was estimated using a Nanodrop One instrument, and an appropriate amount of DNA and sequencing primer was sent for amplicon sequencing (GATC/Eurofins). All amplicons were sequenced in both directions (*i.e.*, with both oMJD135 and oMJD136).

2.2.11 – Chemical transformation of E. coli

Chemically-competent *E. coli* were either purchased from New England Biolabs (NEB, #C2987I), or cultures of *E. coli* were treated with calcium chloride ($CaCl_2$) to produce stocks of competent cells. Chemically-competent cells were produced from 100 ml cultures of *E. coli* which harboured no plasmids, cultured in LB media in a baffled 500 ml Erlenmeyer flask (220 rpm, 37 ˚C) for 5 hours. The culture was then split into 2 x 50 ml Falcon tubes, and cells were pelleted by centrifugation (3,900 x *g*, 10 min, 4 ˚C). The supernatant was discarded and pellets immediately resuspended in 40 ml 0.1 M $CaCl_2$ solution. These suspensions were incubated on ice for 30 min and were then centrifuged (3,900 x *g*, 10 min, 4 ˚C). The cell pellets were washed twice in 20 ml 0.1 M $CaCl_2$, centrifuging after both washes (3,900 x *g*, 10 min, 4 ˚C). Both pellets were resuspended in 2 ml of 0.1 M $CaCl_2$ + 25% v/v glycerol, and this suspension was aliquoted into ice-cold 1.5 ml Eppendorf tubes (50 µl) before being frozen at -80 ˚C immediately.

Chemically-competent *E. coli* were transformed by defrosting an aliquot of competent cells on ice. Plasmid DNA (≤ 5 µl) was added to these cells, mixed briefly and gently using the pipette tip used to add the DNA, and the cells were incubated on ice for ≥ 30 min. Cells were then heat-shocked in a water bath at 42 ˚C for exactly 30 s, and then returned to ice for 5 min. LB broth pre-warmed to 37 ˚C was added to each tube of heat-shocked cells, and the mixture was transferred to a disposable 30 ml Sterilin tube to be incubated at 37 ˚C (1 hr, 220 rpm). From this culture, 150 µl was spread onto LB media containing appropriate antibiotics to select for transformants. An appropriate positive control (1 µl of a stock of either pUC19 or pACYC184) and negative control (5 µl of buffer EB) was included in every transformation experiment. Inoculated plates were incubated overnight and putative clones were sub-cultured for further analysis.

2.2.12 – Molecular cloning of $bla_{\text{CARB-like}}$

To clone $bla_{CARB\text{-}like}$, primers oMJD96 and oMJD97 were used to amplify $bla_{CARB\text{-}like}$ from gDNA extracted from NCTC 30. The primers were designed to incorporate restriction enzyme sites and STOP codons as outlined in Figure 4.22. PCR was carried out using Phusion® high-fidelity DNA polymerase according to the manufacturer's instructions, and a 10 mM dNTP mix (Thermo Scientific, #R0191). Twenty-nine PCR cycles were performed (annealing temperature: 65 ˚C; extension time: 90 s). Reaction intermediates were purified using the QIAquick PCR Purification kit (Qiagen, #28104) according to the manufacturer's instructions.

Approximately 1 µg of both this insert and pACYC184 were digested using 20 units each of BamHI and SalI (NEB, #R3136S and #R3138S respectively) in CutSmart buffer for 15 min at 37 ˚C. One unit of rSAP (NEB, #M0371S) was added to the pACYC184 digest mixture in order to dephosphorylate the 5' ends of the digested vector. Both digestion reactions were continued for a further 30 min before digested DNA was purified as above.

Digested insert and vector were mixed in a molar ratio of approximately 3:1, and ligated using T4 DNA ligase (NEB, #M0202S) at room temperature for 150 min. A ligase-minus reaction was included as a negative control. Ligase activity was inhibited by heating the mixtures to 65 ˚C for 10 min. These ligation mixtures (5 µl) were then used to transform competent *E. coli*.

E. coli that exhibited resistance to both ampicillin and chloramphenicol upon transformation, and were also subsequently confirmed to be tetracycline-sensitive, were cultured and stored as glycerol stocks. Plasmids were prepared from these transformants using the QIAprep Spin Miniprep kit (Qiagen, #27104) according to the manufacturer's instructions. The presence of $bla_{CARB\text{-}like}$ in pMJD61 was checked by PCR using oMJD88 and oMJD89 (homologous to $bla_{CARB\text{-}like}$), and confirmed by Sanger sequencing (GATC/Eurofins) with oMJD98 and oMJD99 (homologous to the sequences flanking *tet* on pACYC184).

2.2.13 – Modified antimicrobial sensitivity assay

A single colony of each strain to be assayed was used to grow a lawn of culture on an LB agar plate overnight (37 ˚C). Plasmid-harbouring strains were cultured on appropriate selective media. Sections of the lawn were suspended in 1.0 ml LB medium and the OD_{600} of this suspension was normalised to 0.5. This suspension was well-vortexed and cotton swabs were used to inoculate an LB agar plate with these standardised suspensions. Plates were allowed to air-dry for 15 min before an MICEvaluator Ampicillin test strip (Oxoid, #MA0110F) was applied to the plate surface. Plates were then incubated for 20 h at 37 ˚C. Break points were determined using the manufacturer's instructions.

2.2.14 – Extraction of plasmids from V. cholerae

Plasmids were extracted from *V. cholerae* using the Qiagen Plasmid Midi kit (# 12143) with minimal modification to the manufacturer's instructions. Briefly, 25 ml cultures of *V. cholerae* were grown overnight at 37 ˚C in baffled 250 ml Erylenmeyer flasks (180 rpm). Cells were collected by centrifugation (3,900 x *g*, 5 min, 4 ˚C) and the supernatant was discarded. The pellet was resuspended in 4 ml buffer P1 containing RNAse A. Buffers P2 and P3 were added according to the manufacturer's instructions, and all samples were kept on ice for ≥15 min before centrifugation (3,900 x *g*, 45 min, 4 ˚C). Debris had not fully pelleted from the samples, and so the supernatant was transferred to a fresh 50 ml Falcon tube and re-centrifuged (3,900 x *g*, 15 min, 4 ˚C). The resultant clarified supernatant was added to a pre-equilibrated Qiagen-Tip 100 and was allowed to flow through the column. The column was then washed (2 x 10 ml Buffer QC) before impure plasmid DNA was eluted in 5 ml Buffer QF. DNA was precipitated using 3.5 ml room-temperature isopropanol and centrifuged (3,900 x *g*, 30 min, 4 ˚C). The supernatant was discarded and the pellet of DNA washed with 2 ml 70% v/v ethanol before

being air-dried and rehydrated with 500 µl buffer TE. These samples were sterility-tested and removed from the CL3 facility.

2.2.15 – Bacterial cultures for transcriptomic experiments

Transcriptomic experiments were carried out using cultures derived from three independent colonies of a bacterial strain of interest. Each strain to be assayed was streaked for single colonies from -80 ˚C stocks onto an LB agar plate, which was incubated overnight at 37 ˚C. Three well-isolated colonies *per* strain were used to inoculate 3 ml overnight cultures in LB media in 30 ml Sterilin tubes, which were grown overnight (37 ˚C, 180 rpm). The OD_{600} of each overnight culture was then determined.

Baffled 250 ml Erlenmeyer flasks were used as the culture vessels for transcriptomic experiments, and contained a pre-prepared culture medium appropriate for the experiment in question (LB media). All media used within an experiment were prepared as part of the same batch, and using the same base reagents, to minimise batch-effect variation amongst transcriptomic experiments. An appropriate volume of overnight culture was added to these flasks to a final OD_{600} of 0.01. The flasks were then loaded into a shaking incubator and grown for an appropriate period of time (either to a target OD_{600} or for a defined period of time post-inoculation). Once the endpoint for the experiment had been reached, the contents of each flask was transferred to a 50 ml Falcon tube and cells were sedimented immediately (3,900 x *g*, 5 min, room temperature). The supernatant was discarded and the resultant cell pellet was immediately frozen at -80 ˚C. All cell pellets pertinent to an experiment were batched together in order to ensure they were all processed on the same day, and under the same conditions, to minimise variability.

2.2.16 – RNA isolation and purification

The Masterpure Complete DNA and RNA Purification kit (Epicentre, #MC85200) was used to extract RNA from *V. cholerae*. Prior to processing samples, MSC surfaces, pipettes, plasticware *etc*., were decontaminated using RNaseZap to remove residual RNase contaminants. Cell pellets were thawed on ice, treated with 300 µl Tissue & Cell Lysis Solution containing Proteinase K, vortexed (10 s) and incubated at 65 ˚C for 20 min with intermittent vortexing. Pellets were cooled on ice. MPC Protein Precipitation Reagent (150 µl) was added

to each sample, tubes were vortexed (10 s) and the contents centrifuged (13,200 x *g*, 10 min, 4 ˚C). If the resultant pellet was unstable, an additional 25 µl MPC reagent was added and the sample re-vortexed and re-centrifuged.

Supernatants were transferred to clean 1.5 ml Eppendorf tubes and total nucleic acids were precipitated using 500 µl room-temperature isopropanol. Tubes were inverted 40 times and then centrifuged to collect precipitated nucleic acids (13,200 x *g*, 10 min, 4 ˚C). Supernatants were discarded and residual isopropanol was removed using a pipette. Each nucleic acid pellet was resuspended in 200 µl DNase buffer containing active DNase I, and incubated at 37 ˚C for 30 min. After this incubation, 200 µl 2 X Tissue and Cell Lysis Solution (TCLS) was added to each sample, followed by 200 µl MPC reagent, to precipitate and inactivate DNase. Samples were vortexed (10 s) and centrifuged (13,200 x *g*, 10 min, 4 ˚C). Supernatants were transferred to clean 1.5 ml Eppendorf tubes and DNA-free RNA was precipitated using 500 µl room-temperature isopropanol. Samples were mixed by inversion 40 times and then centrifuged (13,200 x *g*, 10 min, 4 ˚C). Precipitated RNA pellets were washed twice using 1 ml freshly-prepared 70% v/v ethanol. The pellets were then dried and resuspended in 40 µl nuclease-free TE buffer. RiboGuard RNase inhibitor (1 µl) was added to each sample, which were then tested for sterility before being immediately stored at -80 ˚C.

2.2.17 – RNA integrity assessment

Once sterile RNA was removed from the CL3 facility, it was only defrosted once before being sent for sequencing, at which point its integrity was tested and it was aliquoted into a 96-well plate for submission for sequencing. To assess RNA integrity, an aliquot of each sample (5 µl) was electrophoresed on a 1% agarose gel, to check for the presence of discrete 23S and 16S rRNA bands. Select samples were also analysed using the Agilent Bioanalyser 2100 and the RNA 6000 Nano kit (Agilent, CA USA).

2.2.18 – rRNA depletion and RNA sequencing

rRNA was depleted and sequencing libraries prepared using the Ribo-Zero rRNA Removal Kit (Illumina) by the WSI Bespoke sequencing teams. RNA libraries were strand-specific and sequenced on a 150 PE Illumina 4000 run, multiplexed at 24 samples *per* lane (aiming to produce a minimum of ten million reads *per* sample).

Chapter 3

Contrasting sources and behaviour of epidemic and endemic *Vibrio cholerae* in the Argentinian cholera epidemic, 1992-1998

Contribution statement

Nick Thomson, with Josefina Campos, supervised the work described in this chapter. My thesis committee - Julian Parkhill, Gordon Dougan, and George Salmond - contributed to the interpretation of these results. The isolate receipt data provided by INEI were collated and digitised principally by Tomás Poklepovich. The strain collection was curated by María Rosa Viñas and the *Enterobacterias* team at INEI-ANLIS "Dr. Carlos G. Malbrán", Buenos Aires, Argentina. DNA was isolated from *V. cholerae* by the *Enterobacterias* team, Daryl Domman and I in Argentina. I extracted DNA from samples at WSI with help from Charlotte Tolley.

I performed the analyses described in this chapter, and prepared all of the figures.

Publication

The results, figures, and text presented in this chapter have been submitted for publication.

Dorman MJ, Domman D, Poklepovich T, Tolley C, Zolezzi G, Kane L, Viñas MR, Panagópulo M, Moroni M, Binsztein N, Caffer MI, Clare S, Dougan G, Salmond GPC, Parkhill J, Campos J, Thomson NR. Genomics of the Argentinian cholera epidemic elucidate the contrasting dynamics of epidemic and endemic *Vibrio cholerae*. Revised and re-submitted (*Nature Communications*).

3.1 – Overview

An overall aim of this thesis is to compare the dynamics of pandemic *V. cholerae* to those non-pandemic *V. cholerae* which incur a lesser burden of disease, with a particular interest in describing variation in these distinct *V. cholerae* across time and geography, and at both molecular and phenotypic levels. As previously discussed (Introduction), although there is a clear difference between *V. cholerae* lineages in their ability to cause epidemics and pandemics, studies of gene content variation and even of SNVs have not yet elucidated the specific genetic determinants responsible for "pandemicity". However, these studies have been hampered by relying on relatively small numbers of laboratory strains, particularly of non-pandemic *V. cholerae*, and by a limited understanding of the population structure of the species as a whole.

One possible way to compare pandemic and non-pandemic *V. cholerae* is to examine differences in their evolutionary dynamics during an outbreak of pandemic cholera. Due to limited numbers of samples included in previous sequencing studies, we understand very little about how epidemic *V. cholerae* evolves once it is introduced into a population or a region, or how it contrasts to contemporaneous non-pandemic *V. cholerae* taken from the same region. The only study to date which has sequenced large numbers of closely geographically-related samples over a relatively short period of time has been performed in Bangladesh [235]. However, Bangladesh is hyper-endemic for cholera, and in that setting, multiple sub-lineages of 7PET were found to co-circulate with one another simultaneously. It is also known that there is background immunity to *V. cholerae* in this endemic setting [361]. It would be potentially misleading to infer from studies in such a setting rules that govern 7PET evolution and behaviour upon its introduction into a naïve population. Therefore, to study the evolutionary dynamics of a single 7PET sub-lineage, it was decided to re-visit the Latin American cholera epidemic of the 1990s.

In 1991, cholera returned to Latin America after an absence of nearly 100 years [173, 362]. Crucially, those epidemics which had occurred prior to 1991 had been associated with pandemics 2 through 5, all of which are believed to be due to Classical *V. cholerae* (Introduction, section 1.3.1.1; [36, 97, 162]). Therefore, the introduction of 7PET into Latin America in 1991 is an excellent example of point-source introduction of a pandemic bacterial clone into what can be considered an immunologically naïve population. Latin America

presented a unique opportunity to understand the longitudinal evolution of pandemic *V. cholerae* upon its introduction into a naïve population.

Epidemic cholera broke out in Peru in January 1991, and subsequently spread to Argentina, with the first cholera cases reported on the 5th of February 1992 in Salta province, close to the Bolivian border [174, 363, 364]. On 6th February 1992, a state of emergency was declared by Carlos Menem, then the Argentinian President, in response to outbreaks of cholera in towns in Salta province [365]. Thereafter, cholera cases were reported annually between 1992 and 1998 in Argentina [186, 366–368], with a cumulative total of 4,281 cases reported to PAHO and WHO for this period [363, 369–372].

The rapidity by which pandemic cholera can be transmitted is illustrated by an outbreak linked to Aerolineas Argentinas flight 386. On 14th February 1992, this flight departed Buenos Aires, Argentina, bound for Los Angeles. Following a stop-over in Lima, Peru [373], the flight landed in Los Angeles, carrying 336 passengers and 20 crew members [373, 374]. By 19th February 1992, *V. cholerae* serogroup O1 had been isolated from five passengers suffering from diarrhoea, and by 21st February, one elderly passenger had died from cholera [374–376]. The Argentinian government insisted that contaminated food taken on board during the stop-over in Lima had been responsible for the outbreak; conversely, the Peruvian government claimed that a passenger beginning their journey in Buenos Aires had been the source of the outbreak [377]. Simultaneously, reports of cholera were received by the California Department of Health Services and the Los Angeles County Department of Health Services [373, 374]. Of the 336 passengers on the flight, 100 were found to be carriers of *V. cholerae*, 75 reported diarrhoea of whom ten were hospitalised, one case of which was fatal [373]. Forty-eight *V. cholerae* isolates were obtained from these passengers, of which 34 were serotype Ogawa and 14 were serotype Inaba [373]. It is from this outbreak that the A1552 laboratory isolate, used for numerous functional studies of *V. cholerae* and the source of an important reference genome sequence used in this chapter, was obtained [378, 379].

Argentina is an ideal and unique setting in which to study the evolution of 7PET during the 1990s, because unlike some other countries in the region, the socioeconomic position of Argentina is thought to have enabled the monitoring and control of the epidemic [168, 380]. Argentina instituted mandatory notification of cholera cases nationwide during 1991 after cholera broke out in Peru [368], and developed public information campaigns which resulted

in a concomitant increase in the rate of diarrhoeal disease reporting [381]. During the epidemic period (1991-1998), all suspected cholera cases were tested in microbiology laboratories nationwide, and putative *V. cholerae* from clinical cases of suspected cholera, as defined by the Argentinian Ministry of Health, were isolated and sent to INEI-ANLIS "Dr. Carlos G. Malbrán" (INEI), the national reference laboratory of Argentina, for further confirmation. The INEI archives contain over 3,500 phenotypically-characterised *V. cholerae* isolates, which may represent over 82% of the WHO-reported cholera cases for the whole country from this 1990s epidemic (an epidemic which caused over 1.2 million disease cases of cholera across Latin America [168]).

In order to study the similarities and differences in evolutionary dynamics between epidemic and endemic *V. cholerae*, an appropriate strain collection is required. Previous genomic work looking across Latin America and including five Argentinian isolates [189, 234] suggested that a single toxigenic *V. cholerae* clone belonging to the LAT-1 sub-lineage of 7PET was responsible for the Argentinian cholera epidemic, related to that which caused outbreaks in Peru [189]. Therefore, we hypothesised that the comprehensive strain collection at INEI, together with the metadata recorded for each isolate, would enable the study of the progression of an epidemic attributed to one discrete introduction of 7PET.

Prior to the implementation of genomic analysis, observations made about cholera in Latin America have been difficult to resolve in the light of the cholera paradigm [29]. As discussed previously (Introduction, section 1.3.1.2), this paradigm suggested that epidemic cholera in Latin America was driven by environmental factors, which created an environment conducive to the expansion of local *V. cholerae* populations. These local *V. cholerae* were proposed to have been the aetiological agent of the 1990s epidemic [29, 382]. Although the role of the environment as a source of epidemic *V. cholerae* continues to be debated today [383, 384], the hypothesis that environmental sources gave rise to the Latin American cholera epidemics in 1991 has been shown to be incompatible with genomic data [189]. The Haitian cholera epidemic which began in 2010 has also been shown not to be derived from an environmental source [26, 189]. In addition, genomic evidence has shown that even toxigenic *V. cholerae* O1 that are local to Argentina [367], and to other Latin American countries, are distinct from the aetiological agent of the seventh pandemic cholera [189].

Previous *V. cholerae* sequencing projects have focused on studying isolates from clinical sources taken during outbreaks, and have therefore been enriched for serogroup O1 and toxigenic *V. cholerae* isolates. Indeed, it was only because of a combination of socioeconomic conditions and the materials preserved by certain laboratories that non-pandemic *V. cholerae* O1 lineages were sequenced in our previous study [189]. However, alongside the mandated collection of clinically-isolated *V. cholerae* O1, INEI also received and stored non-O1 *V. cholerae* of both clinical and environmental origin. Crucially, these non-O1 *V. cholerae* were isolated from the same times and places as the *V. cholerae* O1 during the epidemic. After the 1990s epidemic ended in Argentina, environmental surveillance projects saw additional non-O1 *V. cholerae* added to the INEI collection. Thus, it was reasoned that this collection of non-O1 isolates may represent the *V. cholerae* diversity present in, and endemic to, Argentina during and after the period of the cholera epidemic.

This thesis chapter describes an analysis of several hundred genomes of spatiotemporally diverse Argentinian 7PET isolates. These genomic data were then linked to related metadata and to historical literature, to describe in detail the dynamics and trajectory of the Argentinian cholera epidemic. Finally, a number of non-7PET *V. cholerae*, isolated during the time of the epidemic, were contrasted with the pandemic 7PET isolates, laying the foundation for more detailed analyses of non-7PET bacteria in subsequent chapters.

3.2 – Specific aims

This chapter aimed to use the INEI collection of *V. cholerae*:

a) To test the hypothesis of whether isolates belonging to the LAT-1 sub-lineage, which was introduced into South America in 1991, were the aetiological agents of epidemic cholera in Argentina between 1992 and 1998.
b) Using sequencing data, to study how epidemic *V. cholerae* evolved genomically over the course of the epidemic (seven years).
c) To characterise the genomes of non-O1 (and, presumably, non-epidemic) *V. cholerae*, and contrast the genomes of these isolates with those of contemporaneously-isolated pandemic *V. cholerae*.

3.3 – Ethical statement

The isolates described herein were acquired by the national reference laboratories at INEI-ANLIS as part of the routine receipt of notifiable pathogens during this epidemic. No patient data, identifiable or otherwise, were made available for the analyses reported in this thesis. Accordingly, ethical approval was not required.

3.4 – Results

3.4.1 – WHO/PAHO records

In order to obtain an initial insight into the dynamics of the Argentinian cholera epidemic, the number of WHO and PAHO-reported cholera cases for the country were plotted (Figure 3.1). These data are cumulative, and report the number of cases of disease *per annum*.

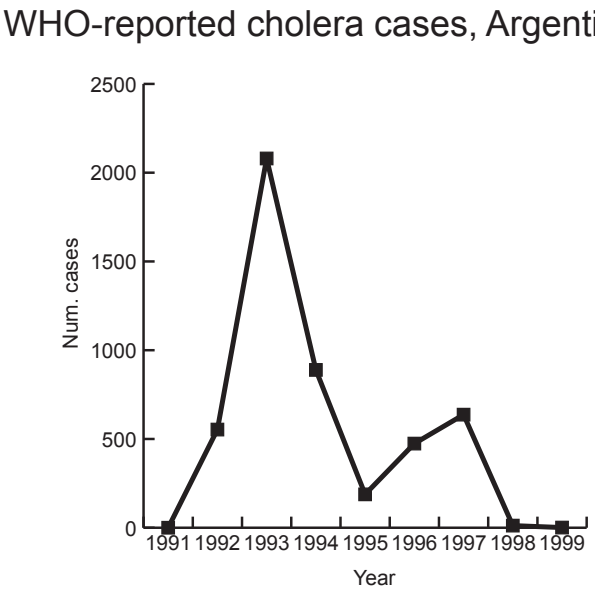


Figure 3.1 – Cholera cases reported to PAHO and WHO from Argentina, 1991-1999. Data were taken from [363, 369–372, 385, 386].

These data suggest that there were two 'peaks' of cholera incidence in Argentina between 1992 and 1998, one in 1993 (2,080 cases) and one in 1996/1997 (474 and 637 cases respectively) (Figure 3.1). However, there are discrepancies between these apparent maxima and those presented in other governmental reports and research publications, which have suggested that there were 'seven epidemics' of cholera in Argentina between 1992 and 1997 [186, 366–368].

3.4.2 – The records of isolate receipt at INEI-ANLIS

All *V. cholerae* received by INEI since 1992 have been recorded and documented, and these records were made available for this PhD thesis. Dates of isolate submission to INEI were available for each isolate, as was the location from which each *V. cholerae* was obtained, the serogroup, and for *V. cholerae* O1, the serotype. GPS co-ordinates were determined for each

province or city from which *V. cholerae* had been received at INEI, and these were used to produce maps of *V. cholerae* O1 and non-O1 receipt over time (Figure 3.2). These were therefore used as a proxy for identifying the location and magnitude of disease incidence across the country (Figure 3.2).

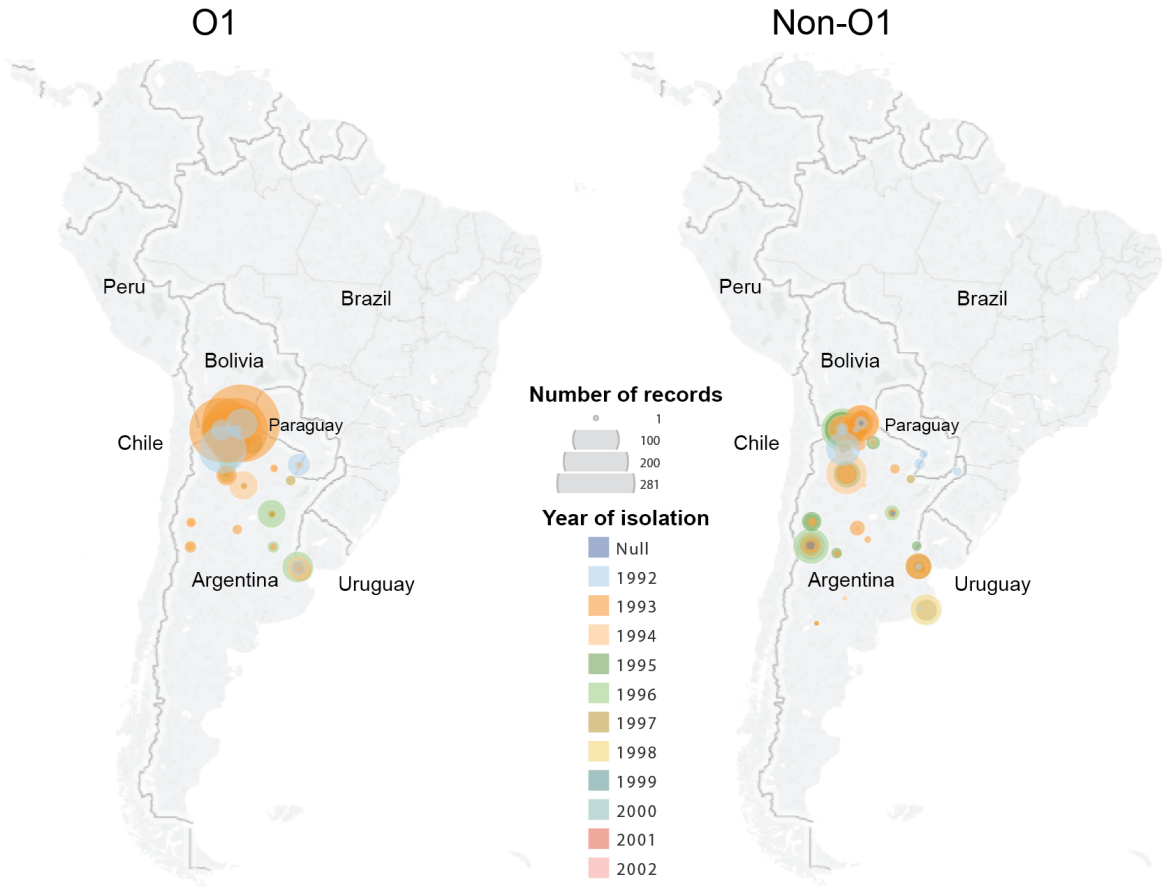


Figure 3.2 – Origins of *V. cholerae* received by INEI, 1992-2002. The non-O1 isolates obtained from coastal regions south of Buenos Aires were obtained from environmental surveillance projects at Instituto Nacional de Investigación y Desarrollo Pesquero (INIDEP), Mar del Plata. The size of the circle scales with the number of isolates received from a particular region in the year indicated.

V. cholerae were received from the North and Centre of Argentina. The geographic area covered by these isolates is approximately 1.2 million km^2 (calculating the area of Argentina North of a latitude line from Buenos Aires). The nature of the national reporting scheme in Argentina at the time, enforced at the national level, made it mandatory that suspected cases of

cholera were tested for *V. cholerae* nationwide. This meant that the absence of isolates from the South represents a true absence of cholera reports and isolation from this region.

The date associated with each bacterial isolate corresponded to the date on which the isolate was acquired by the reporting laboratories nationwide, rather than the date on which the isolate was received by INEI in Buenos Aires. The precise date of isolation was recorded for each isolate. Accordingly, it was reasoned that these data might describe the temporal variation in *V. cholerae* reported in Argentina in greater detail than the PAHO/WHO statistics (Figure 3.1). Dates of isolation were plotted across the period of the epidemic and beyond, to account for environmental sampling that continued beyond the end of the epidemic in 1998 (Figure 3.3).

Of the isolates received, 2,189 were serogroup O1 (60.2 %), and 1,308 were non-O1 *V. cholerae*. Clinical isolates made up the majority (2,077, 94.8%) of *V. cholerae* O1; 112 isolates were either environmental isolates or their sources were not recorded. Of the non-O1 isolates, 714 were of clinical origin (54.5%). There were 134 isolates for which there were no serogroup data recorded (n = 129), that were autoagglutinable (n = 4), or were recorded as being of serogroup O139 (n = 1). Of these 134 isolates, 106 were isolated in January 1993 and correspond to a reported outbreak of *V. cholerae* O1 (Josefina Campos, personal communication). Serotype data were not recorded for 25 of the *V. cholerae* O1 in these records (1.1%).

At least six peaks of *V. cholerae* O1 receipt can be observed in Figure 3.3. A seventh peak may be visible in early 1998. These are consistent with reports of 'seven epidemics' of cholera in Argentina during the 1990s [186, 366–368]. Although there were periods during which no *V. cholerae* O1 were received by INEI (Figure 3.3), non-O1 *V. cholerae* were received more consistently during the 1990s, though their receipt rose coincidentally with peaks in *V. cholerae* O1 receipt. The apparent discordance between these data and the PAHO/WHO records likely reflects the fact that the PAHO/WHO data are only available as annual case/fatality statistics, and are not broken down by month.

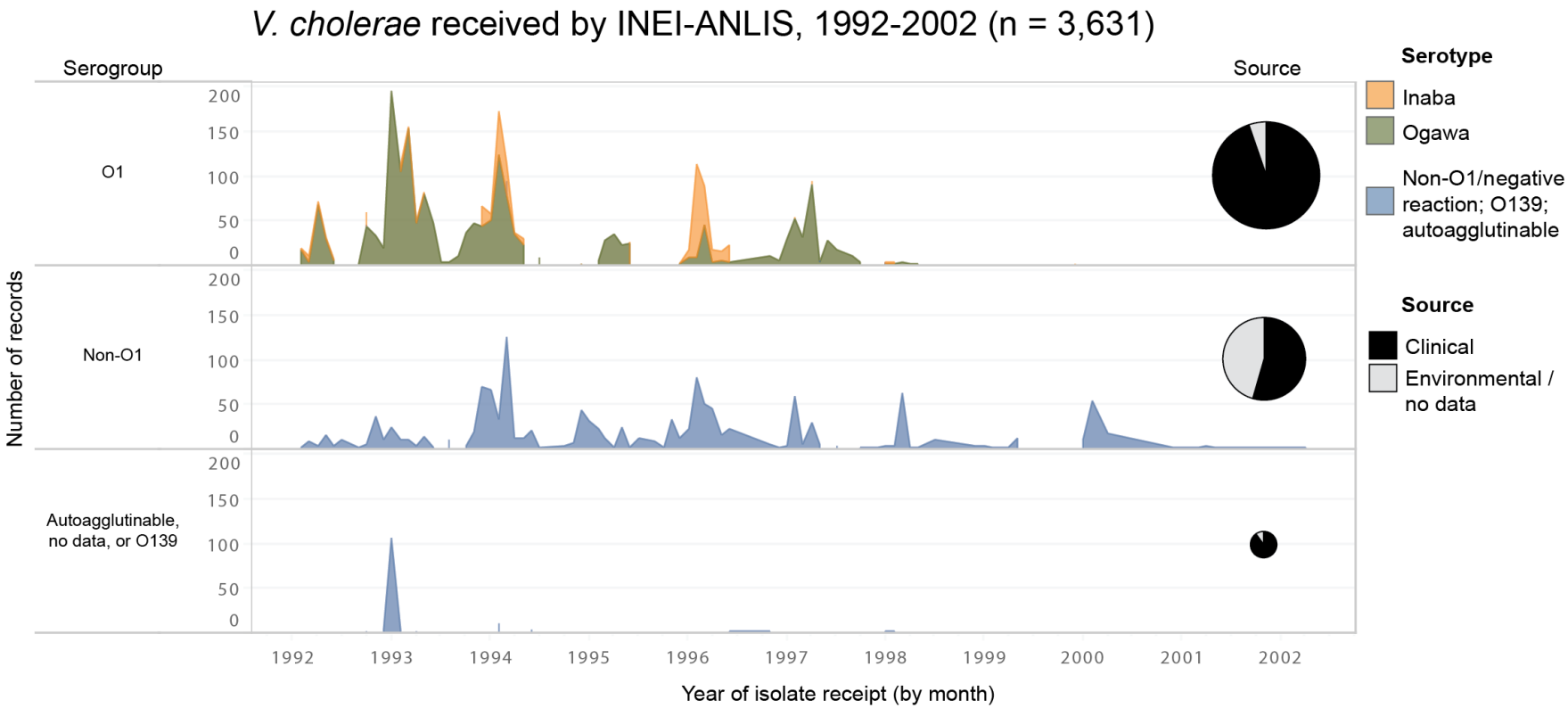


Figure 3.3 – Dates of isolation for *V. cholerae* received by INEI, 1992-2002. X-axis intervals are months of isolation. *V. cholerae* O1 are coloured by serotype. The fraction of clinical to environmental isolates are presented in pie charts (exact numbers are reported in section 3.4.2).

Consistent with a previous report that *V. cholerae* Ogawa was predominant in Argentina during the 1990s epidemic [387], Ogawa isolates accounted for 82.0% of the *V. cholerae* O1 in these records (1,795 isolates), and 369 isolates were serotype Inaba (16.8%). This is despite the initial 1991 Latin American cholera epidemic being ascribed to *V. cholerae* Inaba [369].

3.4.3 – Selection of isolates to sequence for this study

The isolates that were used for this project had been selected in order to represent as much of the country as possible (Figure 3.4) and to span the beginning and the end of the epidemic (Figure 3.5). A strong bias was applied towards sequencing *V. cholerae* O1, from 1992 and 1996/7 and across Argentina, in order to allow questions to be asked about how epidemic *V. cholerae* evolved across space and time during the epidemic. Contemporaneous non-O1 *V. cholerae* from the same geographical regions were also sequenced, in order to compare and contrast the genomes of these non-O1 (and presumably, non-epidemic) bacteria with those of O1 isolates.

Figure 3.4 – Locations from which the isolates analysed in this study were obtained. Of the 490 analysed isolates, 475 had region of origin recorded (96.9%). These data are not stratified by serogroup.

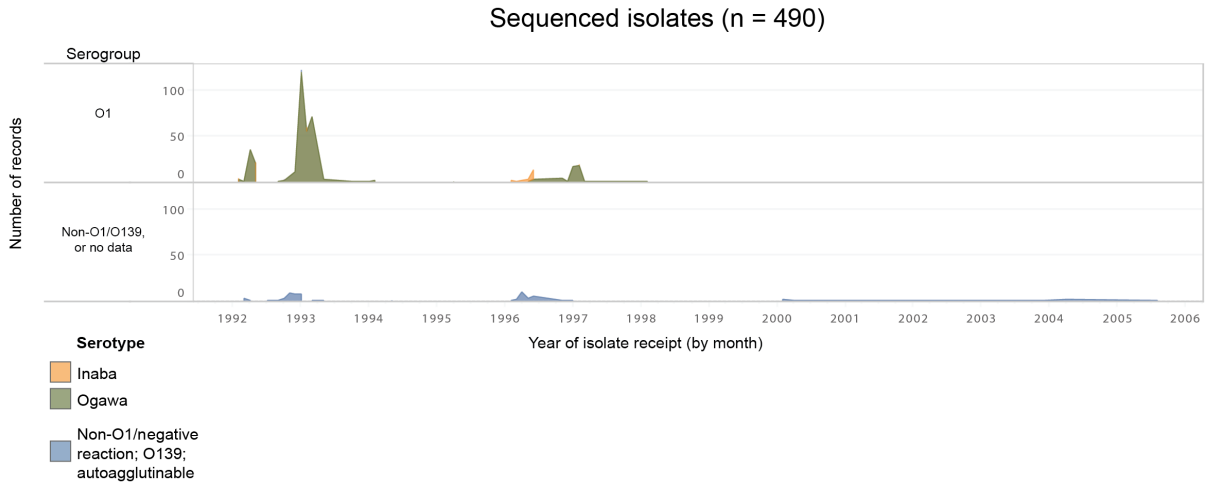


Figure 3.5 – Dates of isolation for the bacteria sequenced and analysed in this study. Dates of isolation were not recorded for four of the sequenced isolates (486/490 isolates represented in this figure; 99.1%).

Genomic DNA was extracted from a total of 511 *V. cholerae* stored at INEI and sequenced on the Illumina X10 platform for the purpose of this PhD project. Twenty-one of the resultant sequences were excluded after failing to meet quality control criteria (see Methods). The distribution of assembly lengths before the application of a 5 Mbp threshold is presented in Figure 3.6 (the *V. cholerae* genome is approximately 4.1 Mbp in length [59]). Kraken reports were also used to identify and remove sequences which were likely to be heavily contaminated with non-*V. cholerae* DNA, or sequences corresponding to species other than *V. cholerae*, as were preliminary phylogenetic trees which were used to exclude isolates on extremely long branches (data not shown).

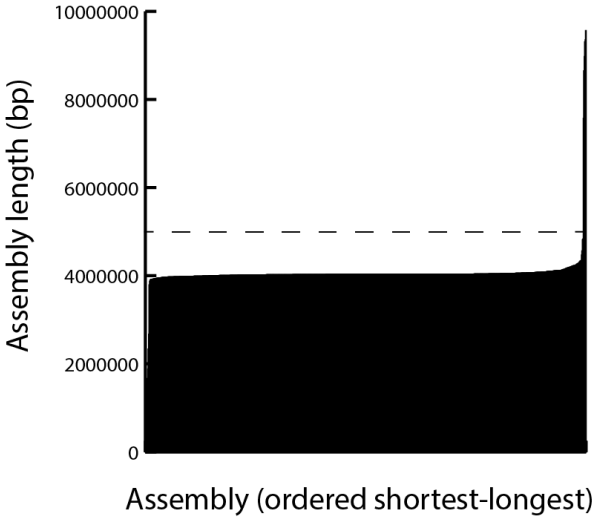


Figure 3.6 – Example of the application of an assembly length cut-off to SPAdes assemblies produced in this study. The dashed line corresponds to an assembly length of 5 Mbp – sequences of a longer assembly were assumed to be contaminated or of poor quality, and were excluded from further analysis. These 936 assemblies include previously-published sequences.

Once low-quality sequences and contaminated samples were excluded, this left a total of 490 genome sequences which were used for all subsequent analysis. A core-gene phylogeny was calculated using the 490 Argentinian sequences, together with a set of 7PET and non-7PET genomes (1,165 total genomes), in order to classify the Argentinian isolates as 7PET, or non-7PET, based on their phylogenetic position (Figure 3.7). Based on these data, 425 of the 490 sequenced *V. cholerae* were determined to be members of 7PET (86.7%) and 65 were classified as non-7PET (Figure 3.7).

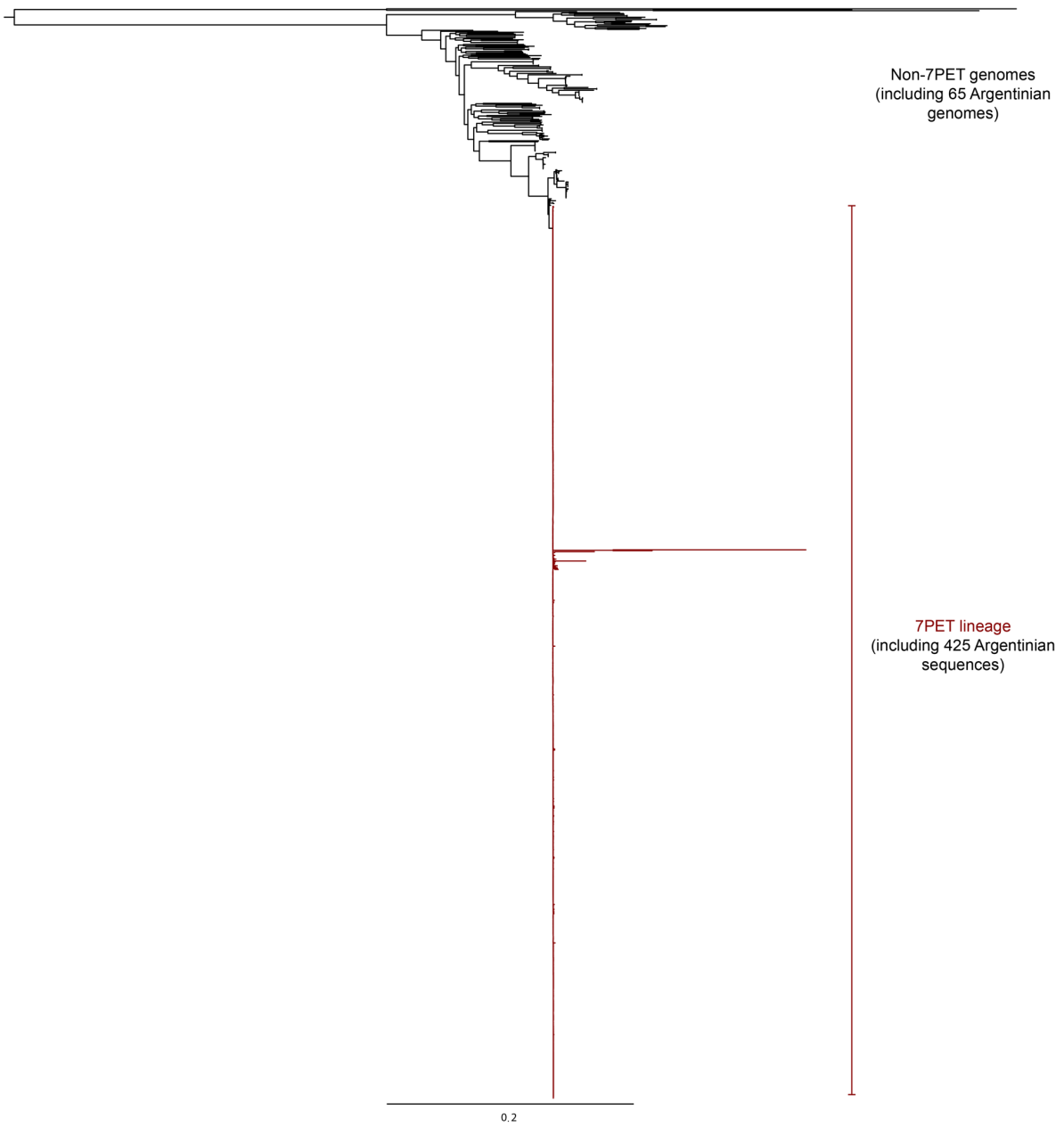


Figure 3.7 – Maximum-likelihood core-gene phylogeny of 1,165 *V. cholerae*. Isolates sequenced in this study (n = 490) were determined to be 7PET, or not, on the basis of their phylogenetic position. The phylogeny was calculated using FastTree v2.1.10 [388] and used as an intermediate analysis step.

3.4.4 – 7PET phylogeny

The 425 7PET genomes were placed into phylogenetic context alongside 517 additional genomes [189], by mapping the reads for these sequences to the *V. cholerae* reference sequence (strain N16961 [59]) and calling non-recombinant SNVs using established analysis pipelines (see Methods). The sequences of both *V. cholerae* chromosomes in the N16961 reference were combined for this purpose. Within the resultant alignment of 942 sequences, Gubbins masked

11.83% of the genome as being potentially recombined. When regions of recombination that were associated solely with pre-pandemic strain M66 were removed from consideration, 2.4% of the genome was predicted to be recombined (Figure 3.8).

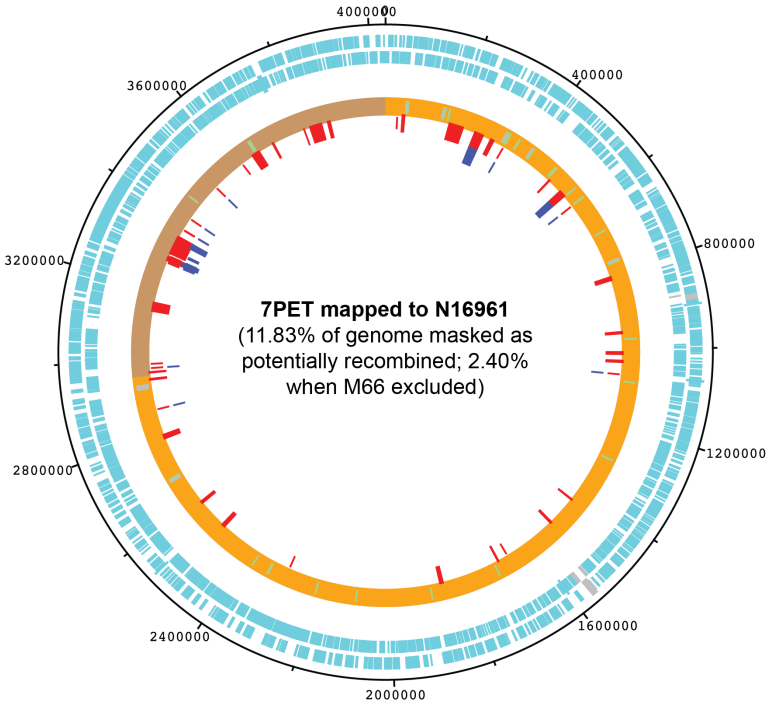


Figure 3.8 – N16961 chromosome regions predicted to be recombined from the 7PET alignment. Regions predicted by Gubbins to be recombined using the alignment of 942 *V. cholerae* sequences are indicated in red. The regions that are not solely associated with the pre-pandemic isolate M66 are indicated in blue (innermost ring). The two outermost rings indicate the presence of open reading frames on the forward and reverse strand, respectively. The sequences of both N16961 chromosomes were concatenated to produce this figure.

A maximum-likelihood phylogenetic tree was then calculated from the non-recombinant SNVs in this alignment (Figure 3.9). Parsimony-informative SNVs were identified from the alignment and used to cluster the data with Fastbaps [334]. These clusters were used to guide the assignment of 7PET isolates to sub-lineages (Figure 3.9). Of the 425 7PET isolates from Argentina, 421 isolates were members of the LAT-1 sub-lineage, which had been introduced to Peru in 1991 and subsequently spread across Latin America [189]. Since none of the Argentinian genomes were members of any other 7PET sub-lineage, the hypothesis that epidemic cholera in Argentina was caused by the same strain introduced into Peru in 1991 is very strongly supported.

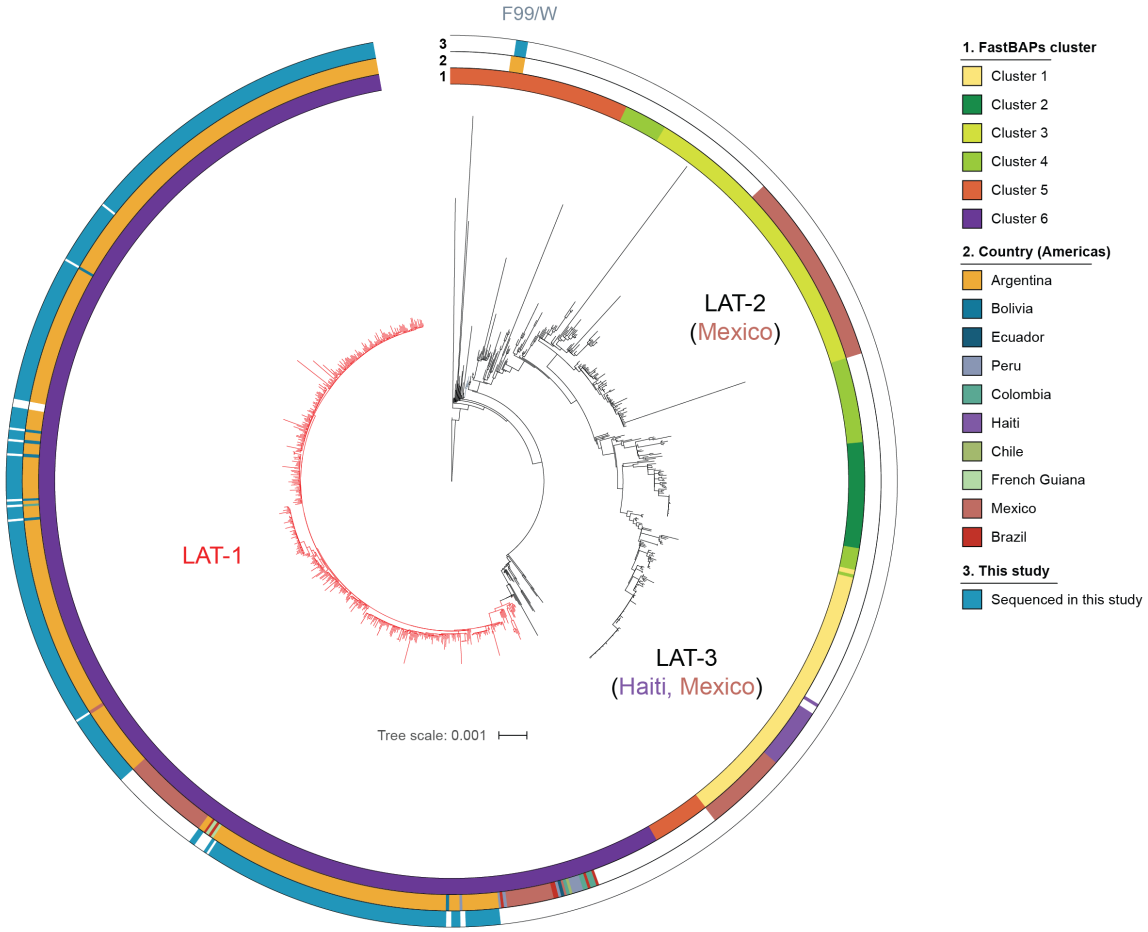


Figure 3.9 – A maximum-likelihood phylogeny of 7PET. Tree calculated using 7,556 non-recombinant SNVs, and rooted on the pre-pandemic isolate M66. Clustering was performed using Fastbaps and an alignment of 3,874 parsimony-informative SNVs. Countries of origin for sequences from South and Central America are reported. LAT transmission events as described previously [189] are indicated; LAT-1 is indicated in red.

The remaining four isolates were not members of LAT-1, but were phylogenetically closely related to F99/W (Figure 3.9). This is a isolate that had been reported previously [189, 389] and had been shown to be neither toxigenic nor a member of LAT-1 [189]. The genomes of the four isolates related to F99/W were investigated in more detail using comparative genomics, to determine the presence and absence of pathogenicity islands associated with 7PET [54, 133] (Table 3.1). All but one were found to encode VPI-1, the genomic island which encodes TCP, the receptor for CTXφ [55]. None of these isolates were found to carry CTXφ or the *ctxAB* genes required to express CT, consistent with the original report of the F99/W genome sequence [189].

Pathogenicity island	F99/W isolate				
	16853_4#67 (F99/W)	26268_1#261 (CCBT0194)	26268_1#259 (CCBT0192)	26268_1#257 (CCBT0190)	26268_1#254 (CCBT0187)
VPI-1	Yes	No	Yes	Yes	Yes
VPI-2	Yes	Yes	Yes	Yes	Yes
VSP-1	Yes	Yes	Yes	Yes	Yes
VSP-2	Yes	Yes	No	Yes	Yes
CTXφ	No	No	No	No	No

Table 3.1 – Presence of select pathogenicity islands in F99/W isolate genomes.

ACT comparisons were used to confirm that these pathogenicity islands are present, and absent, in their totality from the canonical integration sites (based on the N16961 reference genome; Figures 3.10 to 3.12). The absence of CTXφ was also verified using pangenome gene presence/absence data in subsequent analyses, to exclude the possibility of the bacteriophage integrating into non-canonical loci. Similarly, in the case of CCBT0194, mapping data (against N16961) were visualised in order to verify the absence of VPI-1, by confirming that these regions in the reference genome were not covered by mapped reads from the relevant sequencing run (Figure 3.11).

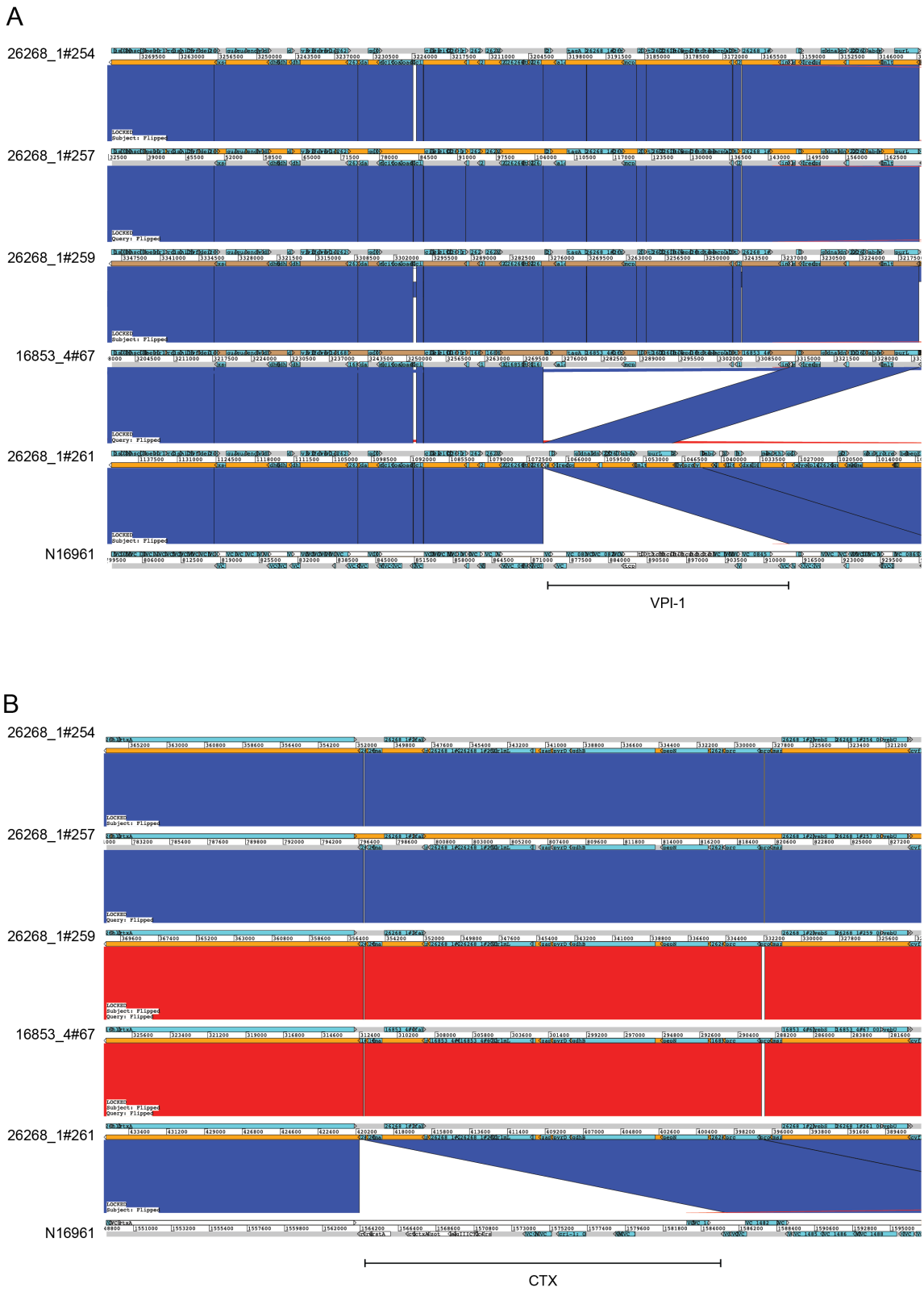


Figure 3.10 – Confirming the presence and absence of VPI-1 and CTXφ in isolates phylogenetically-related to F99/W. (A): The sequence corresponding to VPI-1 is absent from CCBT0194 (26268_1#261) but present in all other sequences being compared, including the N16961 reference. (B): The CTXφ prophage is absent from the canonical integration locus on the larger chromosome in each of the isolates in the F99/W clade (Figure 3.9).

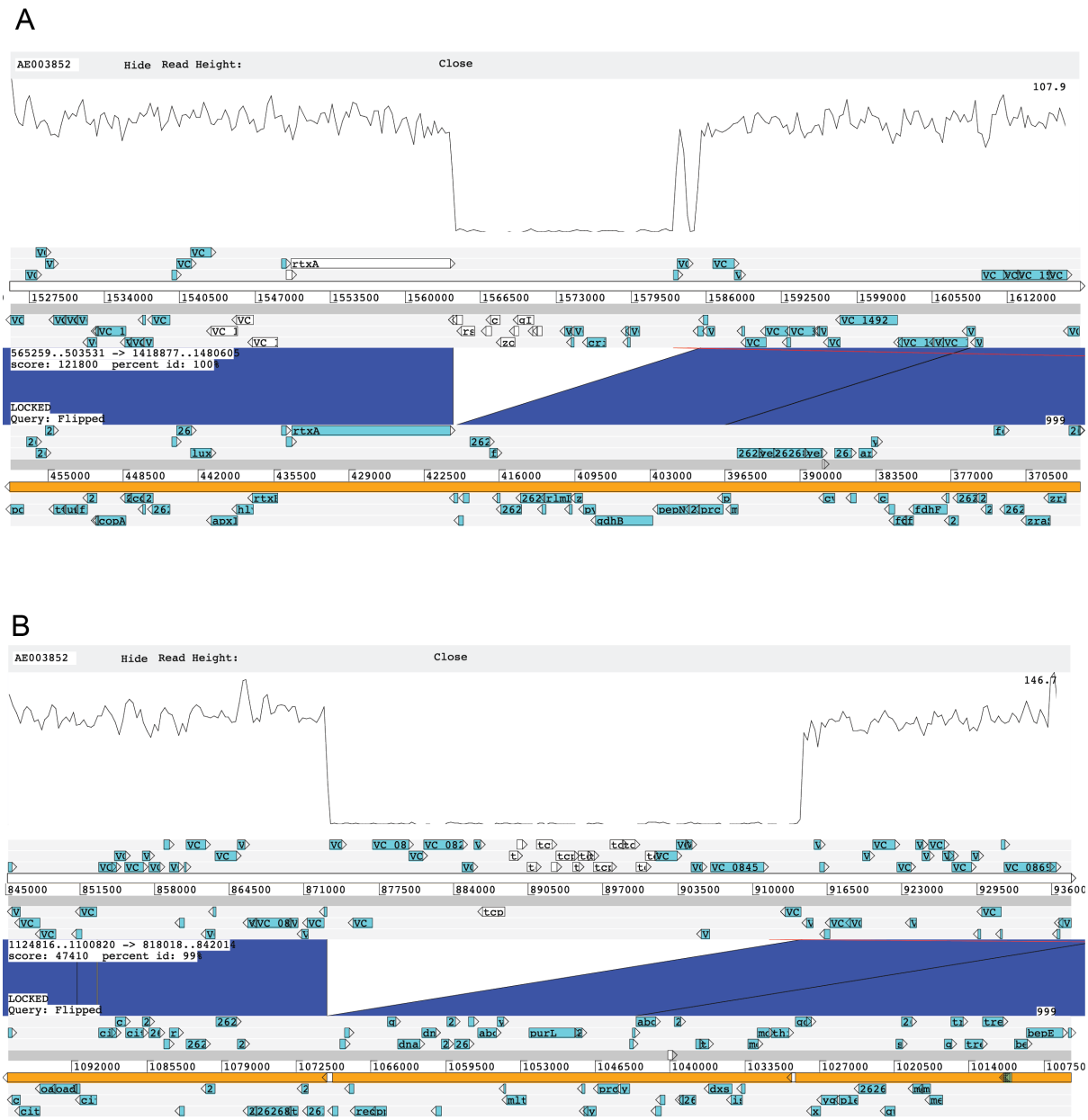


Figure 3.11 – Confirming the absence of CTXφ and VPI-1 from CCBT0194. The depth of reads mapped to the N16961 reference from isolate CCBT0194, the N16961 sequence, and the CCBT0194 assembly are presented. Panel (A) shows the section of the chromosome that harbours the CTXφ prophage and panel (B) shows the VPI-1 integration locus. The drop in coverage to near-zero over these pathogenicity islands, together with the absence of these islands from the CCBT0194 assembly, supports the conclusion that this isolate lacks both CTXφ and VPI-1.

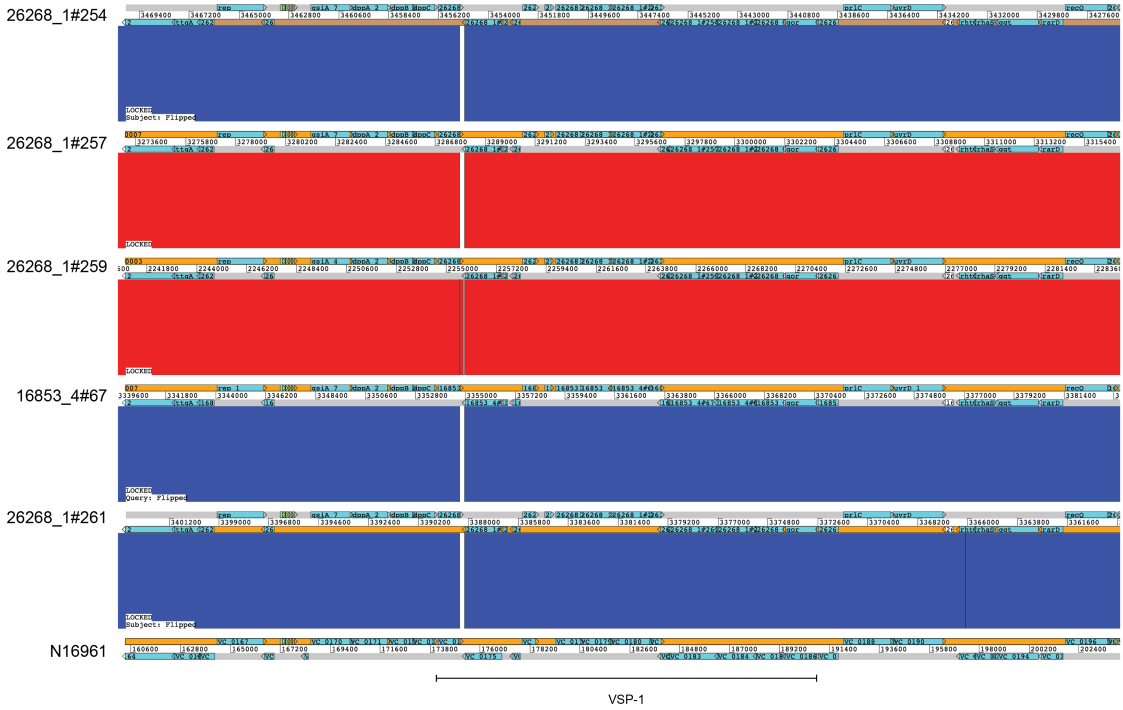


Figure 3.12 – Confirming the presence of VSP-1 in F99/W isolate genomes. VSP-1 is present in its canonical integration locus in all isolates in the F99/W clade (Figure 3.9).

Since the F99/W clade of isolates harbour chromosomal pathogenicity islands that are associated with pandemic *V. cholerae* (Table 3.1), including VPI-1, it is interesting to speculate as to whether these have the potential to be infected by the CTXφ bacteriophage and to become toxigenic. However, live cultures of these isolates were not available with which to test this hypothesis.

3.4.5 – LAT-1 phylogenetics

Although all of the Argentinian LAT-1 genomes clustered together, there appeared to be additional topological structure within the LAT-1 phylogeny (Figure 3.9). Therefore, to study this in greater detail, we re-mapped the reads for the 531 sequences in the LAT-1 sub-lineage to the closed genome sequence of strain A1552 [326], and called SNVs against this reference. A1552 was obtained by the Schoolnik laboratory from the Californian Department of Health Services [379, 390]. It is an Inaba El Tor bacterium [326, 378], and was isolated from a traveller

suffering from cholera in California in 1992 who arrived on Aerolineas Argentinas flight 386 (section 3.1; [378, 379]).

Based on this alignment of 532 sequences, just 0.03% of the A1552 genome was predicted to be recombined (Figure 3.13). A maximum-likelihood phylogeny was calculated using an alignment of 2,651 non-recombinant SNVs (Figure 3.14A). A mean of only 26.05 non-recombinant SNVs across both chromosomes separated the sequence of each LAT-1 isolate from that of the A1552 reference genome (min 10, max 149, stdev 14.10; Figure 3.14B).

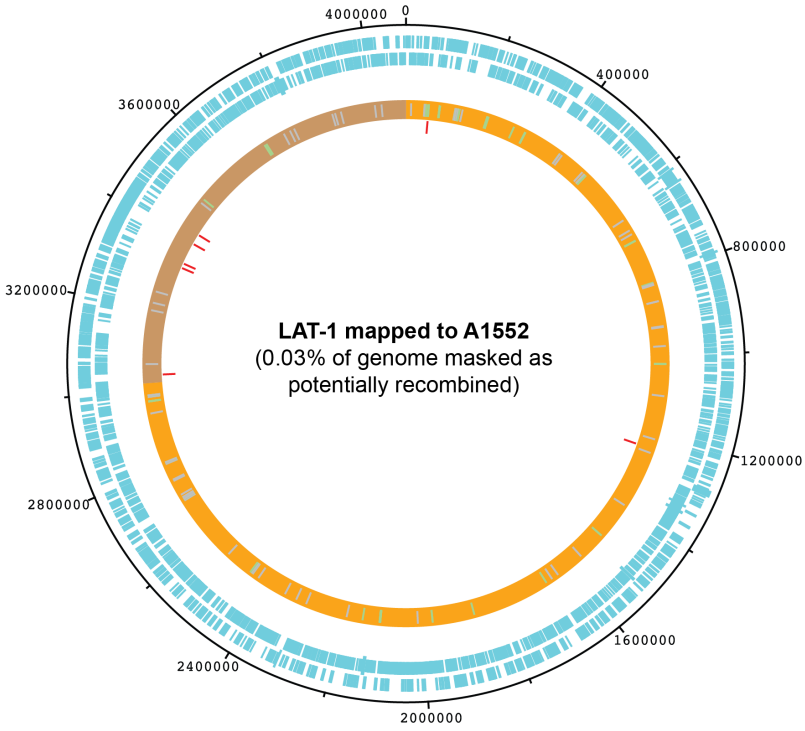


Figure 3.13 – A1552 genome regions predicted to be recombined using the LAT-1 dataset. The sequences of both the larger and smaller chromosomes were concatenated to produce this figure (orange, brown respectively). The inner ring (red) indicates the sections of the genome sequence predicted to be recombined by Gubbins. tRNAs are indicated by green ticks. The two outermost rings indicate the presence of open reading frames on the forward and reverse strand, respectively.

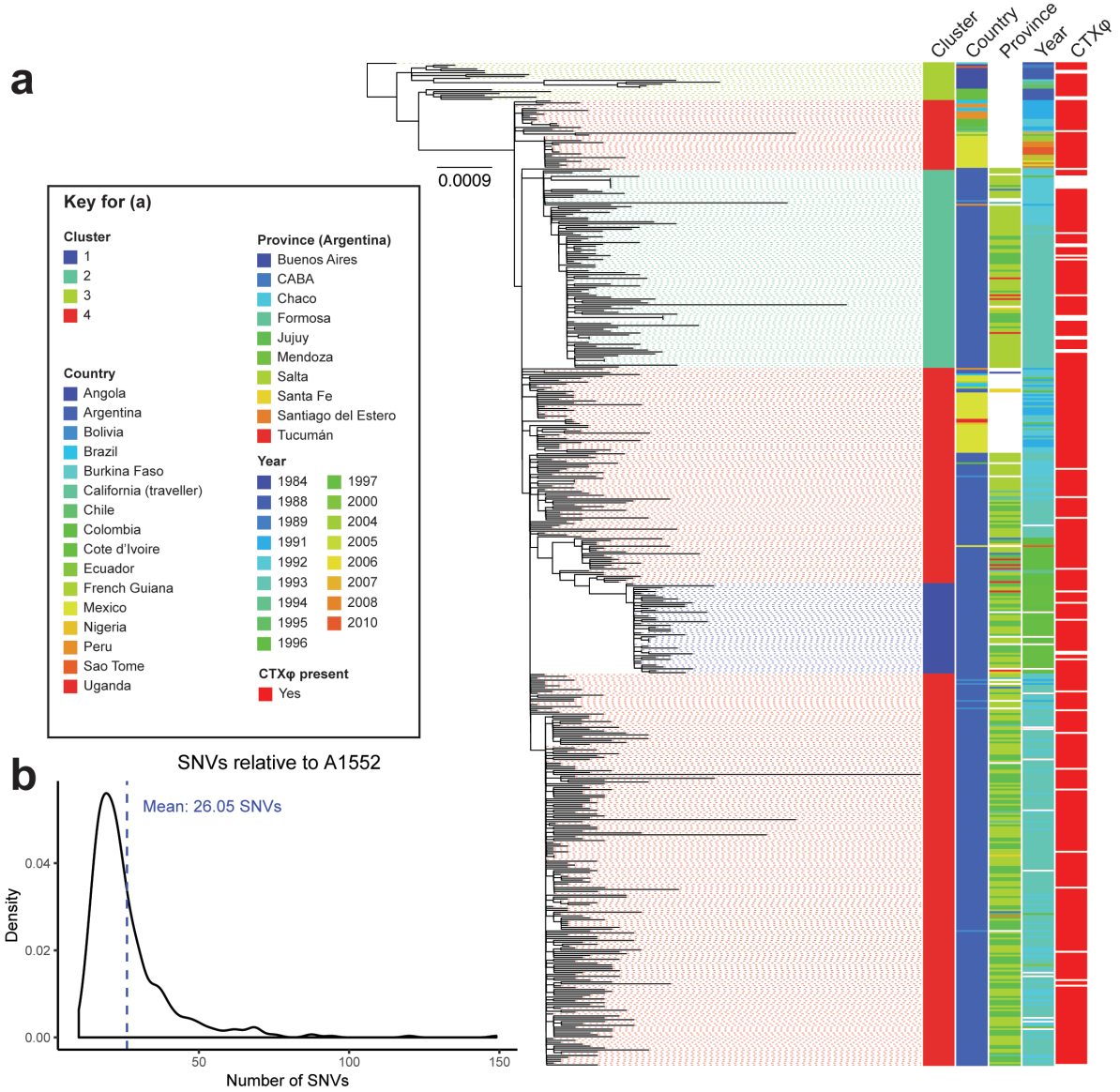


Figure 3.14 – LAT-1 phylogeny and SNV distances from A1552. (A): A maximum-likelihood phylogeny calculated from an alignment of 2,651 non-recombinant SNVs relative to the A1552 reference (accession # CP025936, CP025937). The tree is rooted using the genome of an isolate from Burkina Faso taken in 1984, CNRVC980048. This isolate lacks the WASA-1 genomic marker of LAT-1 [158, 189, 234]. Metadata and the presence of the CTXφ prophage are indicated. An alignment of 725 parsimony-informative non-private SNVs was used to cluster the data using Fastbaps. (B): A density plot of non-recombinant SNVs separating each genome from A1552.

From Figure 3.14 it is clear that LAT-1 sequences did not cluster by geography, either by province or region (Figure 3.14A). Isolates from different Northern provinces were interspersed amongst one another, as were isolates from countries bordering Northern Argentina such as Bolivia. However, limited clustering by date of isolation could be observed. Argentinian isolates from 1996 and 1997 clustered together, and cluster 1 was almost fully

made up of isolates from 1997. Cluster 2 contains Argentinian isolates from 1992 and 1993, and one isolate from 1997, from multiple provinces. Additionally, cluster 2 contained one Bolivian genome from 1992 and a Peruvian genome from 1991.

Consideration was given to whether a dated phylogeny could be used to study transmission of LAT-1 within Argentina. However, a robust temporal signal was not detected in these data – plotting root-to-tip divergence of the LAT-1 phylogeny (Figure 3.14A) against time using TempEst v1.5.1 yielded a regression with a poor correlation coefficient and R^2 value (Figure 3.15). It was also noted that of the 2,651 SNVs in the LAT-1 alignment, 1,926 (72.6%) were private to single genomes in the dataset. This strongly suggested that each isolate may have evolved privately during storage. This may be a consequence of the way in which the isolates were stored; some isolates were stored for up to 27 years before being sequenced, and many had been preserved on stabs or at ambient temperature – as has been described previously, these are storage conditions that can select for hypermutator phenotypes, which would be consistent with accelerated private mutation [308, 391]. Therefore, to avoid drawing conclusions based on potentially spurious results, it was decided that it was inappropriate to calculate dated phylogenies using these data.

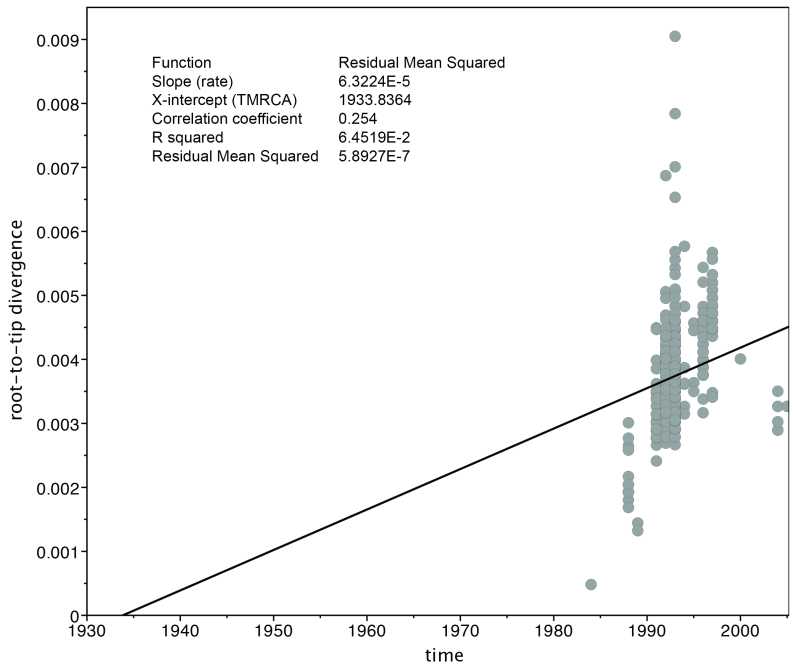


Figure 3.15 – Root-to-tip distance *versus* time for the LAT-1 phylogeny. Argentinian sequences for which dates of isolation were not recorded (n = 4) were arbitrarily assigned to the year 1993. Figure produced using TempEst v1.5.1.

A pangenome was calculated using these 532 LAT-1 genomes. A total of 3,412 core genes were identified from a total of 6,860 genes in the pangenome (core: 97% <= strains <= 100%, see Methods). The gene presence/absence matrix was interrogated to determine the presence and absence of WASA-1, *ctxAB*, and CTXφ across the dataset, and to detect genes associated with IncA/C plasmids (Figure 3.16).

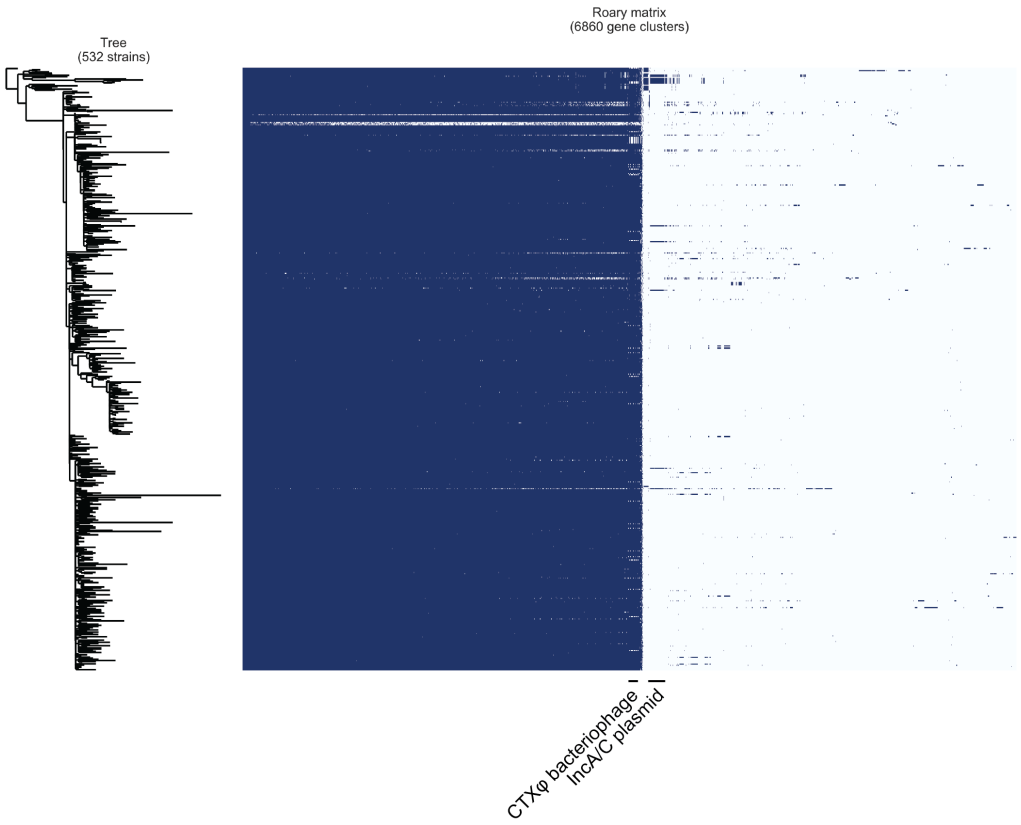


Figure 3.16 – LAT-1 pangenome gene presence/absence matrix visualisation. Phylogeny is as presented in Figure 3.14A. The position of CTXφ bacteriophage genes and those genes present on IncA/C plasmids are highlighted. White horizontal lines in the core genome block correspond to a small number of poorly-assembled sequences, which have been included to contextualise these sequences with previously-published studies.

In addition to the small accessory genome in these LAT-1 isolates, it was observed that the addition of new genomes to this analysis contributed negligible numbers of new genes to the pangenome overall – after the addition of 532 sequences, just ~1,800 unique genes had been added (Figure 3.17). This contrasts strongly with that observed in the non-LAT-1 genomes (see section 3.4.10).

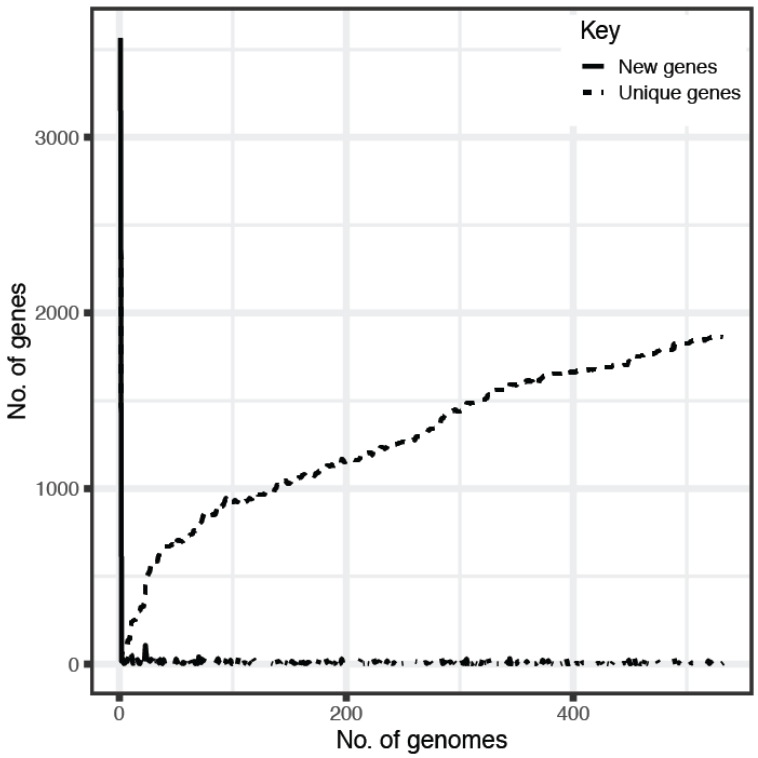


Figure 3.17 – Gene discovery as LAT-1 genomes are added to the pangenome. The initial peak of new gene detection corresponds to the identification of the core genome.

The presence of WASA-1 within LAT-1, a genomic island serving as a marker of the lineage [234], was confirmed by extracting the complete nucleotide sequence of WASA-1 from the A1552 genome sequence and using it to query each LAT-1 assembly with BLASTn. In order to confirm the presence of *ctxAB* and CTXφ, and to detect antimicrobial resistance genes and plasmid replicons, ARIBA was used together with the ResFinder and PlasmidFinder databases, and a custom database of *V. cholerae* virulence genes (Figure 3.18). ARIBA was also used to identify variation in the sequence of the *wbeT* gene, mutations in which are responsible for the switch from Ogawa to Inaba serotype [197, 198] (Introduction, section 1.3.1.3).

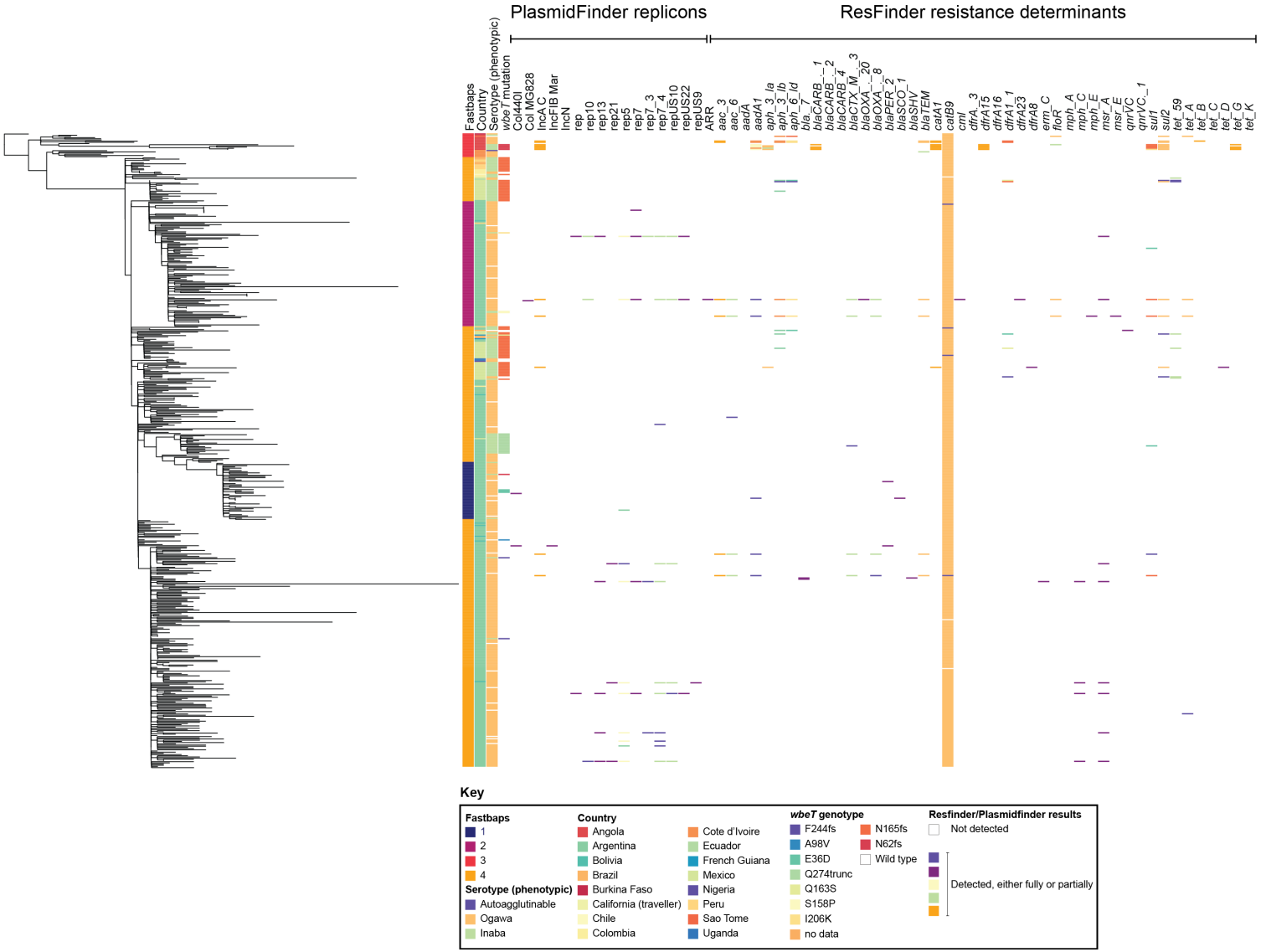


Figure 3.18 – Antimicrobial resistance genes, plasmid replicons, and *wbeT* genotype variants within LAT-1. Data generated using ARIBA (Methods, section 2.1.13).

3.4.6 – Inaba and Ogawa serotype variation within LAT-1

The serotype of *V. cholerae* O1 is of particular historical importance in Latin America. It is known that the initial 1991 cholera epidemics in Peru and elsewhere in Latin America were associated with serotype Inaba *V. cholerae*, which became dominated by serotype Ogawa bacteria in 1992 and thereafter [208, 369]. The Argentinian epidemic began in 1992 and was reportedly dominated by *V. cholerae* Ogawa [387], despite the initial cholera epidemic in Peru being ascribed to *V. cholerae* Inaba [369]. INEI records support this, and also suggest that the dominant serotype in Argentina varied between Inaba and Ogawa over time (Figure 3.3). This suggested subtleties to the dynamics of Ogawa and Inaba serotype *V. cholerae* O1 in Argentina during the epidemic, and begged the question of whether outbreaks due to Inaba bacteria in Argentina after initial Ogawa outbreaks was due to an Inaba strain being imported from elsewhere, or to seroconversion within the lineage already present in Argentina.

As described in the Introduction (section 1.3.1.3), methylation of the terminal sugar on the O1 lipopolysaccharide chain by the WbeT enzyme confers an Ogawa serotype on *V. cholerae* O1, and if the *wbeT* gene is disrupted, WbeT activity is abolished, and an Inaba phenotype results [197–199]. Therefore, to determine a genetic explanation for the apparent flux in serotype in Argentina over the course of the epidemic (Figure 3.3), the genotype of *wbeT* was determined for every isolate in LAT-1. Within the sub-lineage, nine distinct *wbeT* mutations were identified which were predicted to disrupt the WbeT protein (Figure 3.19). It was assumed that mutations that were predicted to frameshift or truncate translated *wbeT* would cause an Inaba phenotype (N62fs, N165fs, F244fs, Q274trunc), as would other mutations that either were found in Inaba isolates in this dataset (I206K) or were otherwise known to confer an Inaba phenotype (S158P [200]). Since all of the isolates harbouring the E36D *wbeT* mutation had an Ogawa phenotype, it was assumed that this mutation does not abolish WbeT function. Of the remaining eight mutations, the genomic predictions correlated well with the phenotypic serotype assigned to each isolate; the *wbeT* genotype matched the phenotypic serotype for all but two of the of the 398 serotyped LAT-1 isolates sequenced in this study (99.4% concordance; Figure 3.19).

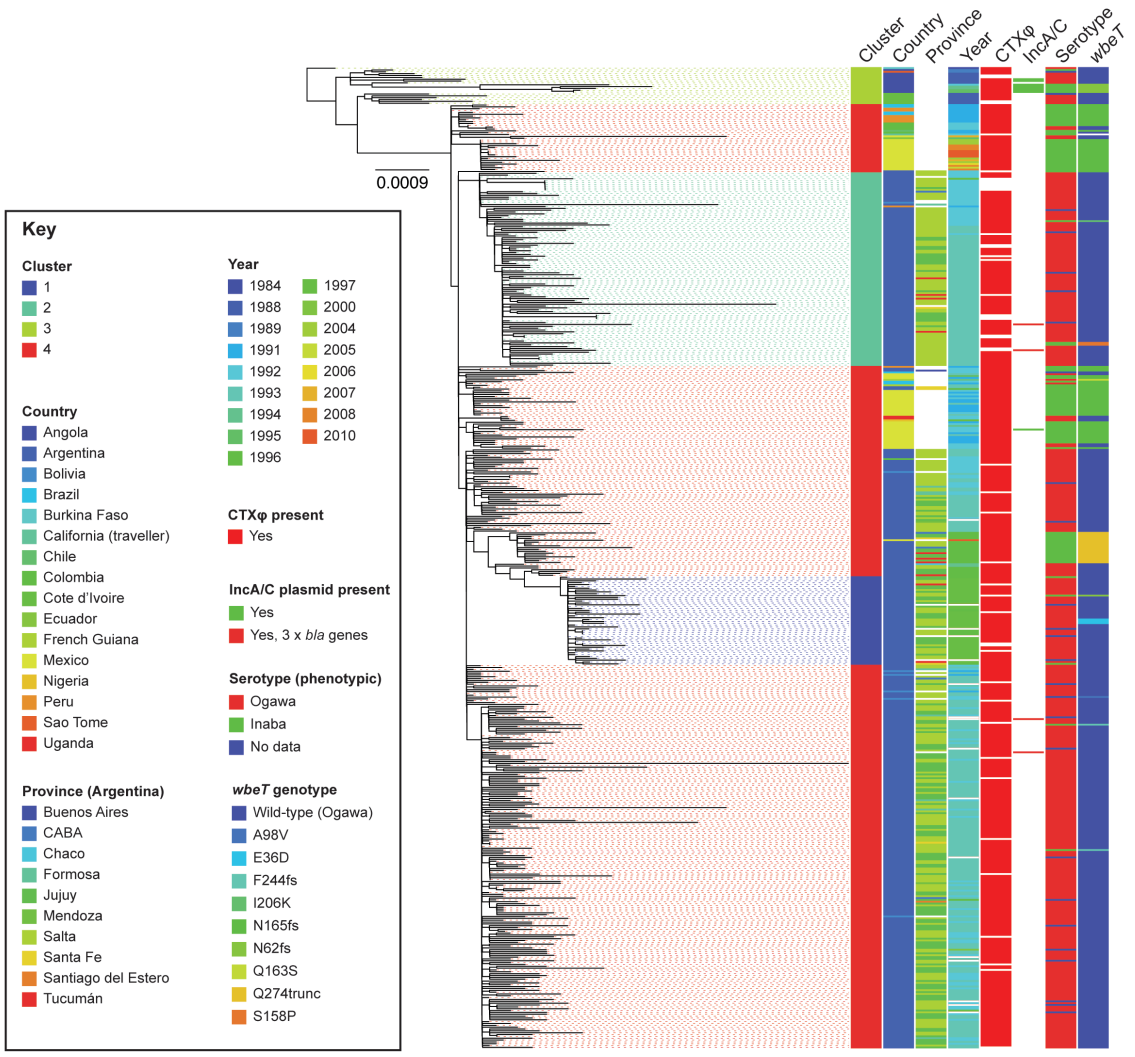


Figure 3.19 – Variation in *wbeT* genotype across the LAT-1 phylogeny. The maximum-likelihood phylogeny presented in Figure 3.14A is re-drawn here with additional metadata and results of select ARIBA analysis results. Missing data are indicated by white space.

Genome data show that the Peruvian Inaba isolates from 1991 harbour the N165fs mutation in *wbeT* (Figure 3.19). Since LAT-1 was introduced into Latin America from West Africa [189], we compared these data to West African Inaba isolates sharing a common ancestor with LAT-1, but post-dating the Peruvian outbreak. These Angolan isolates were collected between 1992 and 1995, just after LAT-1 had been introduced into Peru, and were found to harbour a different mutation, N62fs (Figure 3.19) [158, 189]. Sixty-eight LAT-1 isolates collected since 1991 harbour the N165fs mutation, including isolates from Brazil, Mexico, Chile, Argentina and Colombia, as well as isolates from Peru, all of which were phenotypically serotype Inaba (Figure 3.19). Environmental isolates from Mexico collected between 2004 and 2010 also harbour this mutation, and are part of the same cluster of isolates (Figure 3.19). Hence, the

N62fs and N165fs mutations are likely to have arisen independently, prior to spreading within West Africa and Latin America, respectively.

It has been hypothesised that cholera entered Argentina through the North of the country, which shares borders with Chile, Bolivia, Paraguay, and Brazil [174]. The LAT-1 phylogeny includes genomes from bacteria collected in 1991 and 1992 from Chile, Bolivia and Brazil [189, 234] (Figure 3.19). These were either serotype Ogawa (Bolivia, n = 7; Brazil, n = 1) or Inaba (N165fs; Brazil, n = 6; Chile, n = 1), and were interspersed amongst contemporaneous serotype Ogawa isolates which were collected in Northern provinces of Argentina (Figure 3.19). All were members of cluster 4, except for one Bolivian genome (1992) which was a member of cluster 2, as discussed previously (section 3.4.5; Figure 3.19). Cluster 4 also contains the A1552 reference sequence, which is of an Inaba genotype (N165fs; Figure 3.19). This observation further supports the hypothesis that the same *V. cholerae* sub-lineage circulated within, and between, countries at the Northern border of Argentina.

Once the concordance between phenotypic serotype and genotypic inference had been established, these data were compared to the longitudinal data detailed in Figure 3.3. Following a relative lull in cholera cases in 1995, cases of cholera resurged in Argentina during 1996 [369]. This was associated with serotype Inaba *V. cholerae* (Figure 3.3). Seventeen Argentinian Inaba *V. cholerae* isolates from 1996 formed a closely-related subclade within cluster 4 of the LAT-1 phylogeny (Figure 3.19). These isolates contain a unique mutation in *wbeT*, Q274trunc, and the subclade includes one 2010 Inaba isolate from Mexico (Figure 3.19). Additionally, this subclade shares a common ancestor with the clade of 48 isolates from 1997, which are serotype Ogawa and comprise cluster 1 (Figure 3.19). The 1996/1997 outbreak was not geographically-restricted; this cluster contained isolates from multiple provinces (Figure 3.19).

3.4.7 – Plasmids and antimicrobial resistance in LAT-1

The identification of 3,368 core genes suggested that ~89% of the 3,776 annotated genes in the A1552 reference genome are core to LAT-1. This also implied that gene gain and loss within LAT-1 was also very rare (Figure 3.16; 3.17). However, some examples of gene gain/loss were identifiable; fifty-one of the genomes in the LAT-1 phylogeny were found to lack the CTXφ prophage in its entirety (Figure 3.16; Figure 3.19). It is possible that this loss was a result of long-term culture as has been noted previously [235, 392]. There was also evidence of sporadic

gene gains. For instance, four Argentinian LAT-1 *V. cholerae* were found to harbour genes encoding the extended-spectrum β-lactamases (ESBLs) $bla_{\text{CTX-M-3}}$, $bla_{\text{OXA-8}}$, and bla_{TEM} (Figure 3.18). Manual interrogation of the assemblies for these isolates confirmed that these three ESBL genes were carried on contigs that also included IncA/C plasmid replicons (Figure 3.20).

IncA/C replicons were also detected in Angolan isolates from 1988 and the early 1990s, consistent with previous reports [158, 189] (Figure 3.18, 3.19). These isolates form part of cluster 3 in this dataset, which includes other African genomes pre-dating the transfer of LAT-1 into Latin America in 1991 (Figure 3.18, 3.19). Although two of the Angolan genomes did harbour bla_{TEM} (Figure 3.18), the complement of resistance determinants in these isolates does not match those found in Argentinian *V. cholerae*.

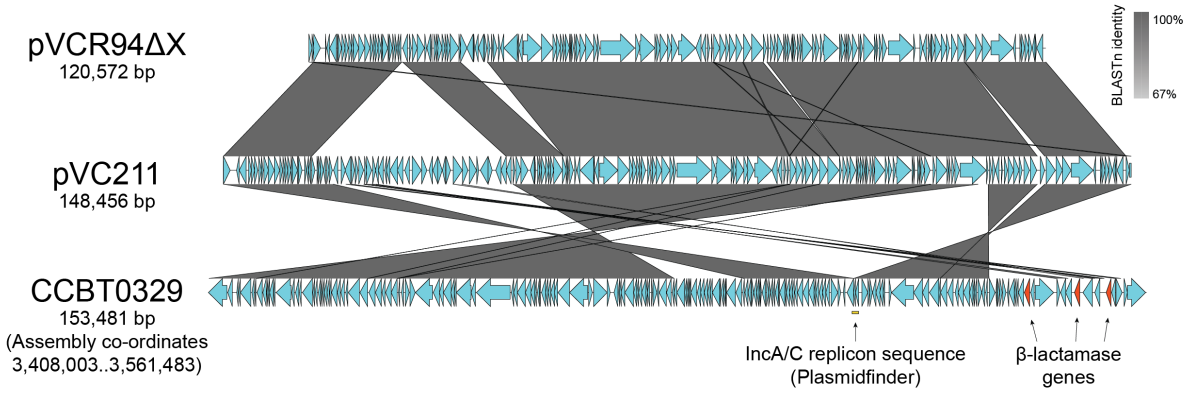


Figure 3.20 – Comparison of a fully-assembled IncA/C2 plasmid from Argentinian isolate CCBT0329 to published *V. cholerae* multidrug resistance plasmids. The positions of the IncA/C replicon sequence present in the PlasmidFinder database and the three genes encoding β-lactamases in the CCBT0329 contig are indicated. The sequences and annotations for pVCR94ΔX and pVC211 were obtained from Genbank (accession # KY399978.1 and KF551948.1, respectively; [393, 394]). The CCBT0329 sequence was annotated with Prokka.

Multidrug resistance plasmids encoding ESBLs have been previously reported in Argentinian *V. cholerae* [366, 387], where molecular characterisations of these large plasmids demonstrated that they can be mobilised and conjugated into *E. coli* [366]. Since IncA/C plasmids are conjugative [395], the plasmids identified in these sequencing data encoding $bla_{\text{CTX-M-3}}$, $bla_{\text{OXA-8}}$, and bla_{TEM} are very likely to correspond to the MDR plasmids seen in *V. cholerae* O1 isolated during the 1990s epidemic from Argentina [366].

3.4.8 – Phylogenetic contextualisation of Argentinian non-7PET isolates

Having performed an in-depth characterisation of LAT-1, and found very limited genetic variation at the levels of SNVs, gene gain and loss, and homologous recombination, attention was turned to the 65 non-7PET Argentinian genomes, to compare the observable dynamics of these isolates to those of LAT-1. In order to contextualise the 65 non-7PET isolates sequenced in this study, these were combined with 318 published non-7PET genome sequences, including three *Vibrio* spp. A pangenome was calculated from these sequences, and 201,790 SNVs extracted from an alignment of 2,719 core gene sequences was used to calculate a maximum-likelihood phylogeny for these isolates (Figure 3.21). The collection of genomes used for contextualisation included a set of recently-reported Chinese genomes [396]; these include isolates which were shown to be closely-related to lineages of *V. cholerae* O1 that had reported to be local to Latin America [189].

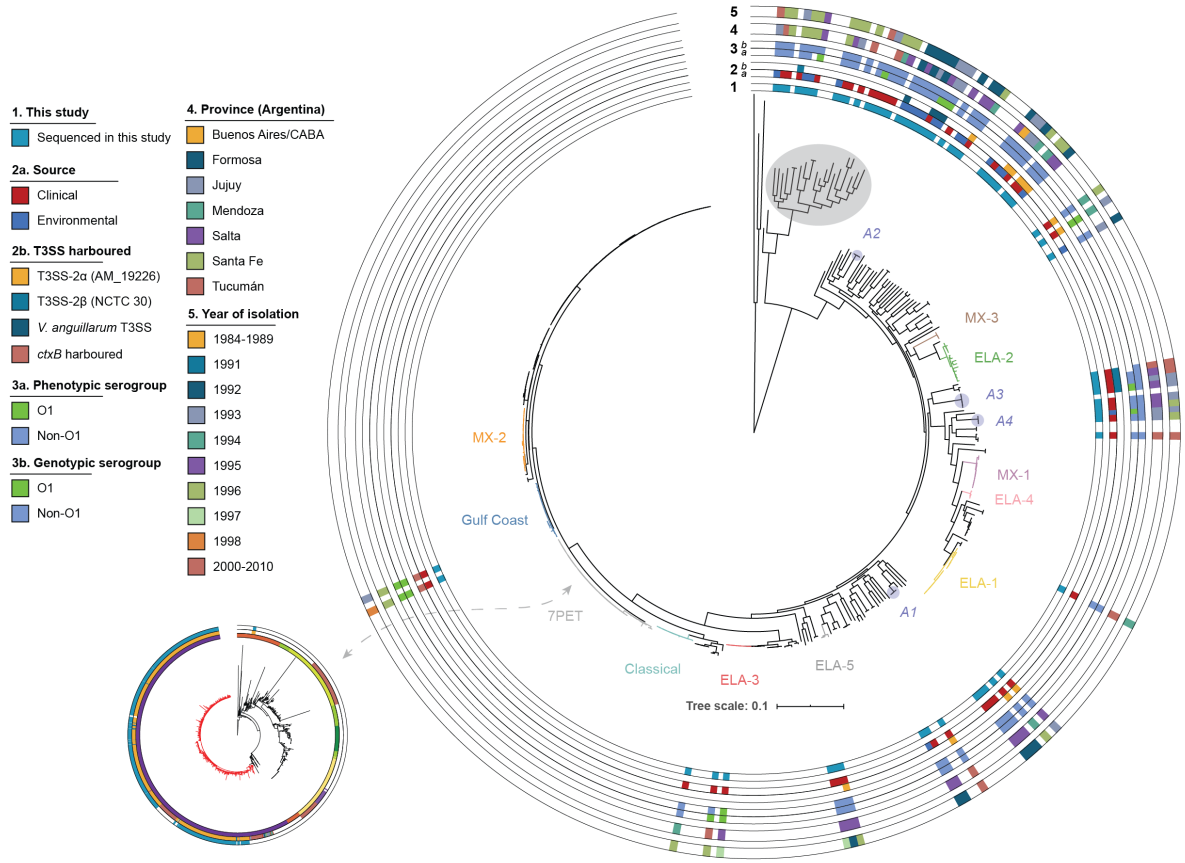


Figure 3.21 – Non-7PET *V. cholerae* phylogeny. A maximum-likelihood phylogenetic tree calculated from 201,790 SNVs identified from an alignment of 2,719 core genes (see Methods). Metadata are presented for the isolates sequenced in this study. The grey disc indicates a cluster of *V. cholerae* which are genetically distinct from pandemic lineages but are still members of the *V. cholerae* species, and will be discussed in greater detail in subsequent chapters. Included in this cluster is a non-toxigenic *V. cholerae* O139 isolate, described in Chapter 4.

In order to confirm the O1 serogroup status of these 65 isolates, the presence of the nucleotide sequence of the serogroup O1 operon extracted from N16961 was determined for each isolate using BLASTn. Four isolates were phenotypically and genotypically serogroup O1. Two of these were members of the previously-described Gulf Coast lineage of *V. cholerae* O1, including the single sequenced *V. cholerae* O1 from 1998 (Figure 3.21). Both Gulf Coast isolates harboured CTXφ and were toxigenic, and the two remaining *V. cholerae* O1 isolates were members of ELA-3 [189] (Figure 3.21). All four isolates were of clinical origin. The remaining 61 isolates lacked the genes required to produce cholera toxin, and were confirmed *in silico* not to harbour genes encoding the O1 antigen, though 45 of these were of clinical origin (Figure 3.21).

Four new lineages of non-O1 non-7PET *V. cholerae* were identified amongst these isolates, defined as clades formed by three or more Argentinian non-7PET isolates in the phylogeny. These were labelled as A1-A4, where A stands for 'Argentina' (Figure 3.21). These lineages contained isolates that were of clinical origin alone (A1, A3) or clinical and environmental origin (A2, A4), were acquired in different years (A3, A4), and from different regions (A2, A3, A4), suggesting that these represent populations of non-7PET *V. cholerae* local to Argentina (Figure 3.21).

3.4.9 – Type III secretion systems in Argentinian non-7PET isolates

Of the 61 non-O1 non-7PET isolates, 21 harboured one of three distinct Type III secretion systems (Figures 3.21, 3.22). These virulence determinants have been associated with pathogenic non-O1 *V. cholerae* as well as with other pathogenic Vibrios [303]. The three T3SSs detected in these isolates included the T3SS-2α described in *V. cholerae* AM_19226, an isolate commonly-used for functional and regulatory studies of T3SS [266, 397]. A less-common system described by Carpenter *et al*., T3SS-2β, was also identified [397], as was a third putative T3SS element which most closely resembles a T3SS detected in two virulent Chilean *Vibrio anguillarum* isolate genomes [398] (Figure 3.22). This putative T3SS was found in lineage A2. The presence of T3SS-2β in lineage A3 was of particular interest – A3 is composed of clinical isolates, contains a previously-described Argentinian isolate, TUC_T2734 [189], and includes one isolate from Salta province collected in the year 2000. T3SS elements were mutually exclusive; no more than one T3SS was detected in a genome. It is also clear from these limited data that more T3SS-positive isolates were of clinical origin

than environmental (T3SS-2α: 10 clinical, 2 environmental; T3SS-2β: 5 clinical, 0 environmental; *V. anguillarum* element; 1 clinical, 3 environmental). No T3SS were detected in 7PET.

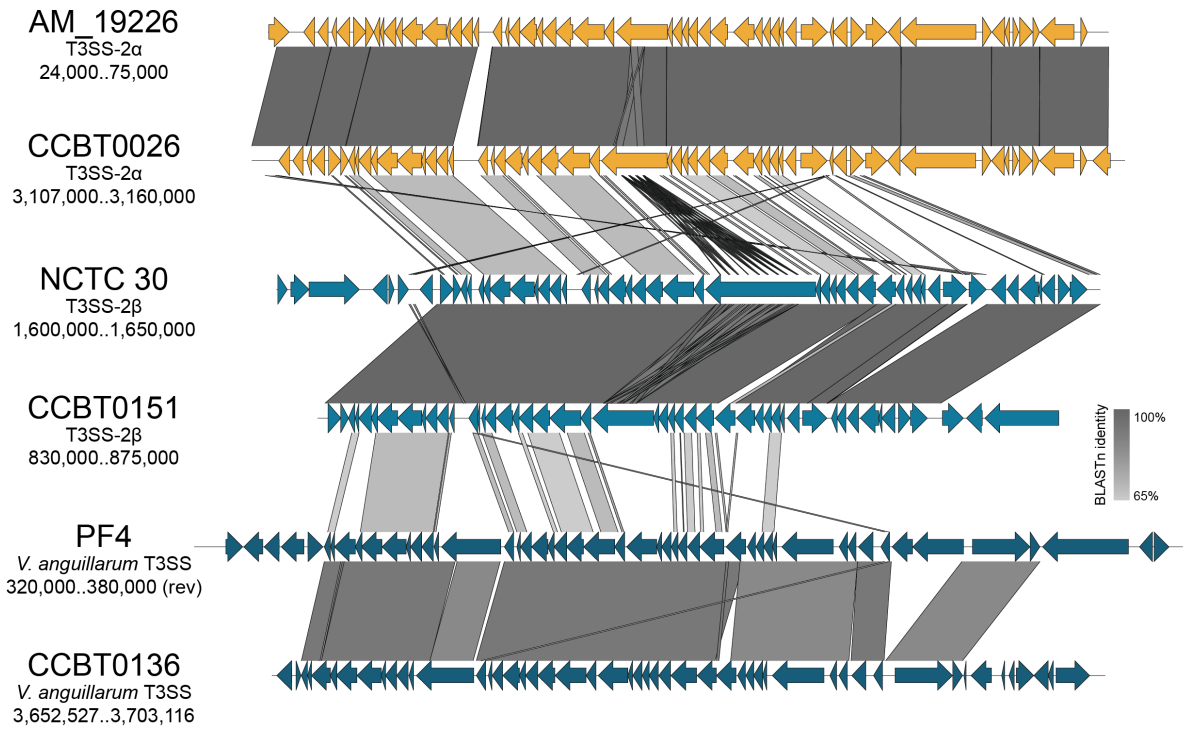


Figure 3.22 – Comparison of T3SS elements detected in Argentinian *V. cholerae* against T3SS taken from reference sequences. Co-ordinates for each region of the assemblies that have been aligned are reported (rev = reverse orientation). Annotations were obtained from Prokka-annotated assemblies used to calculate the pangenome, or from Genbank (PF4 and NCTC 30; accession # CP010081.1 and LS997867.1, respectively). The presence of T3SS-2β in NCTC 30 is discussed in Chapter 4 (section 4.3.10). The PF4 sequence was reflected manually to produce this figure (Adobe Illustrator).

3.4.10 – Comparison of Argentinian LAT-1 and non-7PET pangenomes

Having assigned all of the 490 Argentinian genomes to 7PET or non-7PET, maps of isolate origins stratified on lineage were produced (Figure 3.23). Although non-7PET isolates were fewer in number than 7PET, they were obtained from the same geographical regions and the same time periods as the 7PET isolates (Figure 3.23). This acted as reassurance that the O1 and non-O1 *V. cholerae* sequenced in this study had sampled the same times and regions as one another, albeit that *V. cholerae* O1 had been sequenced in larger numbers than non-O1 (Figure 3.5). This also means that, in principle, both contemporaneous 7PET and non-7PET bacteria should have opportunity to access similar gene pools and to participate in HGT.

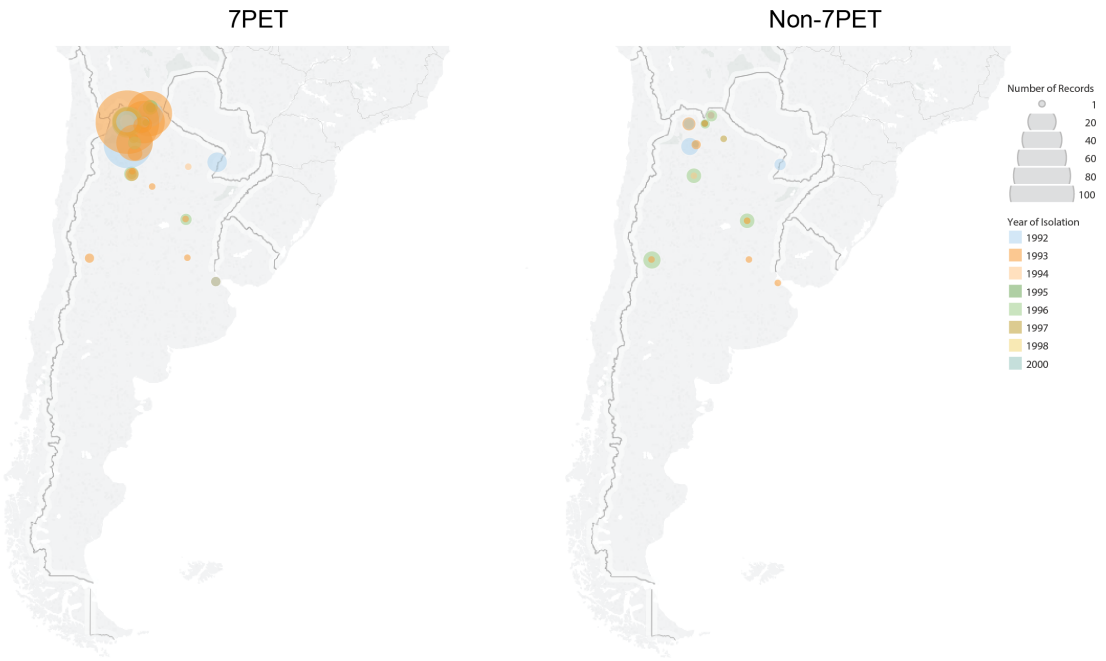


Figure 3.23 – Map showing geographic origin for isolates used in this analysis. Data from Figure 3.4 were stratified by lineage. The size of each circle scales with the number of isolates assigned to that year. Circles coloured by year.

The non-7PET isolates had a considerably expanded accessory genome when compared to LAT-1 (23,458 cloud genes in the collection of 383 diverse genomes compared to 3,313 in the 532 LAT-1 genomes) (Figure 3.24). Thus, the ratio of core genes to cloud genes in LAT-1 was approximately 1:1, whereas in the non-7PET pangenome, the core:cloud gene ratio was ~1:8.

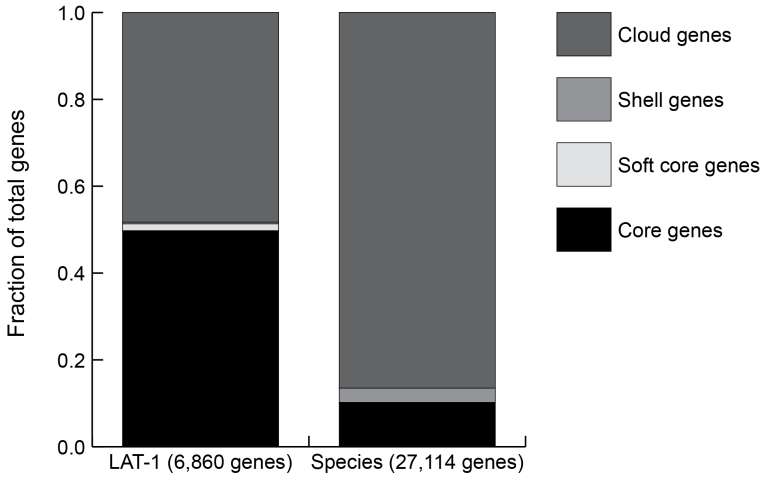


Figure 3.24 – Comparing the summary statistics for the LAT-1 and *V. cholerae* species pangenomes calculated in this study. Fractions of the total number of genes in the pangenome are presented. Definitions used: core: 97% <= strains <= 100%; soft core: 95% <= strains < 97%; shell: 15% <= strains < 95%; cloud: 0% <= strains < 15%.

The large non-core genome was also evident when the pangenome gene presence/absence matrix was visualised against the *V. cholerae* species phylogeny (Figure 3.25). This contrasted starkly with that of the LAT-1 dataset (Figure 3.16).

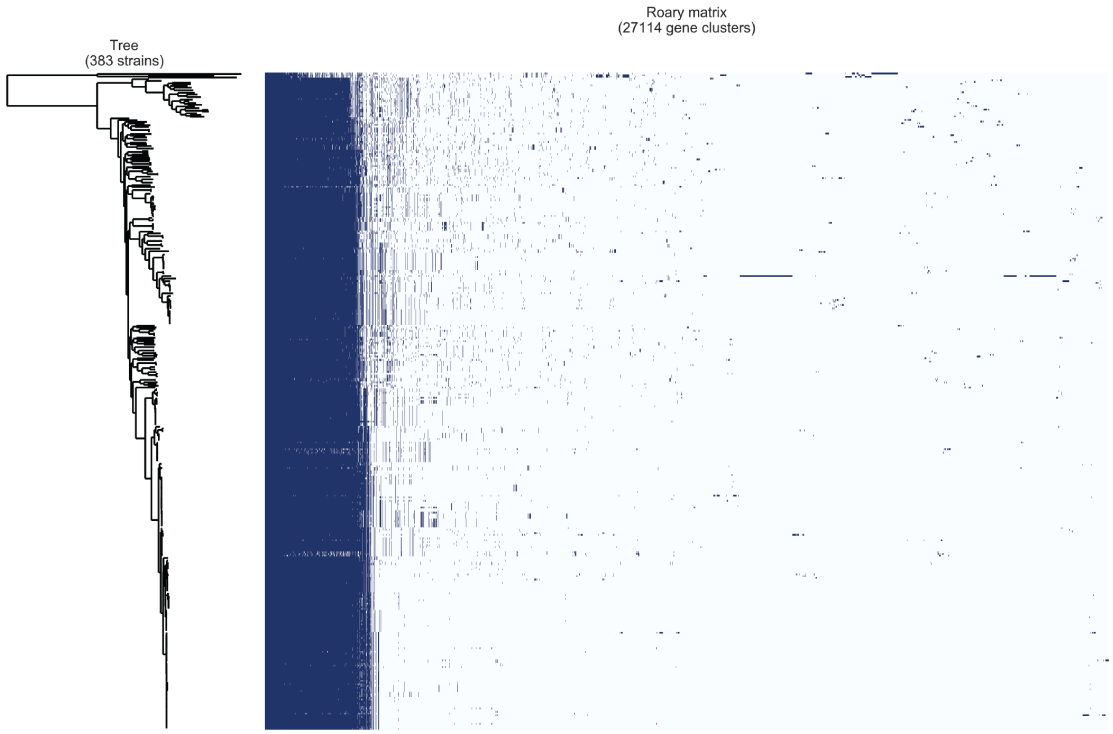


Figure 3.25 – Visualisation of the *V. cholerae* pangenome gene presence/absence matrix. Phylogeny is as presented in Figure 3.21 and is similarly rooted on three *Vibrio* spp. genomes.

The rate of gene discovery as sequences were added to the non-7PET pangenome (Figure 3.26) was much greater than was observed in the LAT-1 pangenome (Figure 3.17), despite there being 38% more sequences in the LAT-1 pangenome. Unlike the LAT-1 pangenome, in which the discovery of new genes was rare, the core of 2,719 genes in this dataset is dwarfed by the number of unique genes discovered as diverse genomes are added to the pangenome. Where the addition of 532 genomes to the LAT-1 pangenome saw ~1,800 unique genes being identified (Figure 3.17), after adding 383 non-7PET genomes to the *V. cholerae* pangenome, ~13,000 unique genes were detected (Figure 3.26).

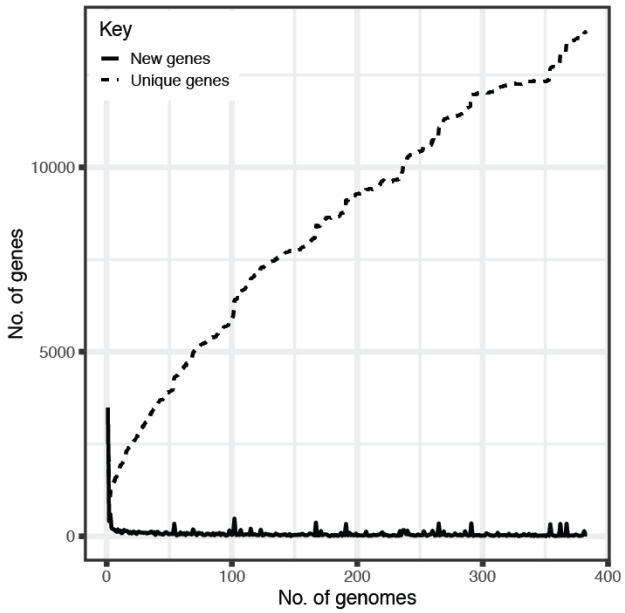


Figure 3.26 – Identification of new genes as genomes are added to the *V. cholerae* pangenome.

The contrast between the clonality of LAT-1 and the diversity of the *V. cholerae* species dataset was also evident when other summary statistics were plotted for these pangenomes (Figure 3.27). Of particular note are the differences in the numbers of total genes relative to conserved genes, the overall contribution to the number of pangenome genes made by adding each additional genome to both datasets, and the number of genes that have a diminishing BLASTp identity in the species pangenome compared to that of LAT-1 (Figure 3.27).

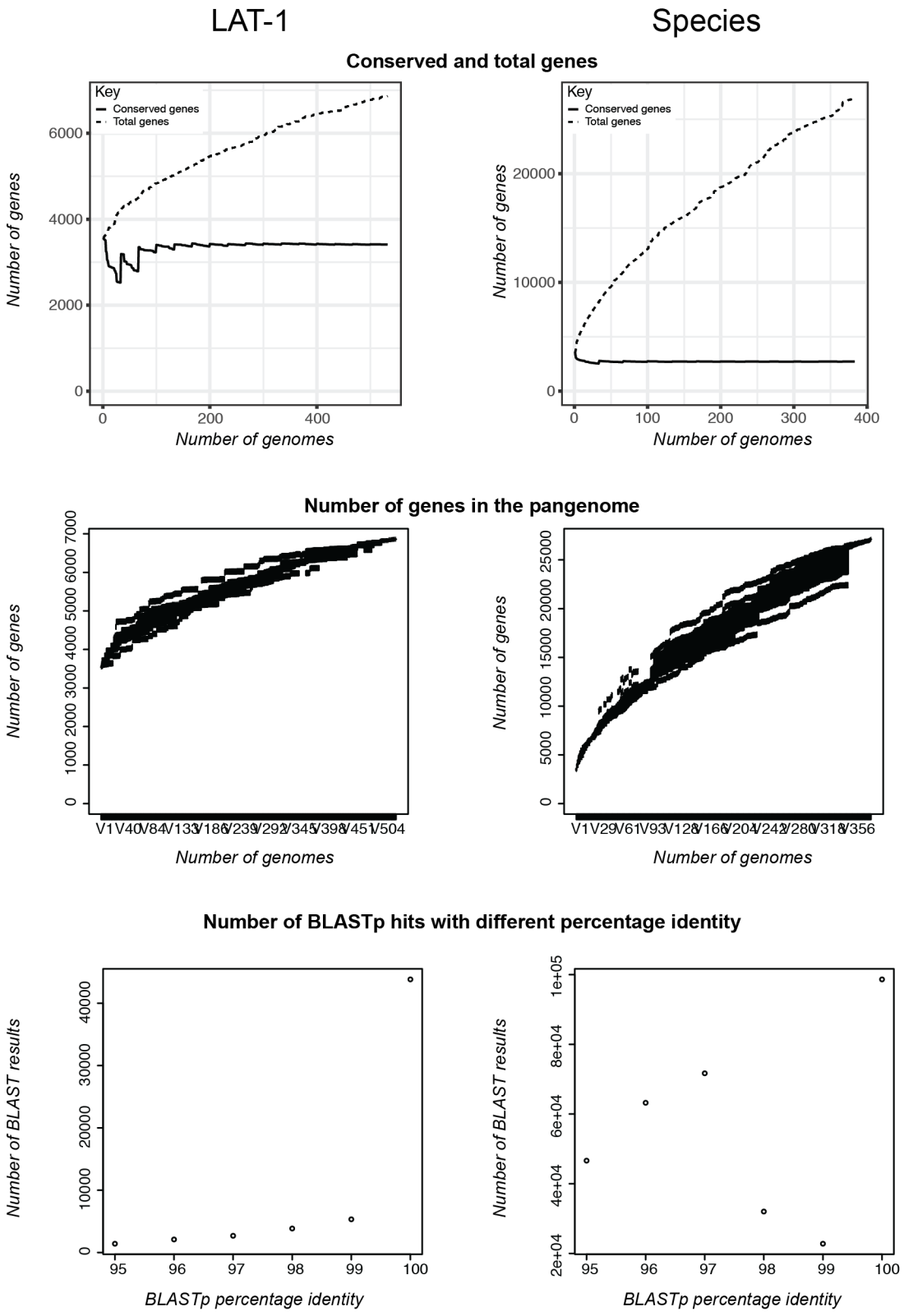


Figure 3.27 – Summary statistics for LAT-1 and non-7PET *V. cholerae* pangenomes.

As another measure of genetic distance, average nucleotide identity (ANI) values relative to the A1552 reference sequence were calculated for all genomes sequenced in this study. By this measure, the non-7PET isolates were also highly genetically diverse in comparison to the 7PET genomes, with a mean ANI relative to A1552 of 97.61 (min 95.90, max 99.65, stdev 0.960; Figure 3.28), in contrast to LAT-1 (mean ANI 99.99, min 99.96, max 99.998, stdev 0.0032; Figure 3.28). This finding is consistent with the phylogenetic analysis presented above (Figure 3.21). It should be noted that an ANI value of 95% is commonly accepted to be threshold for separating species [340].

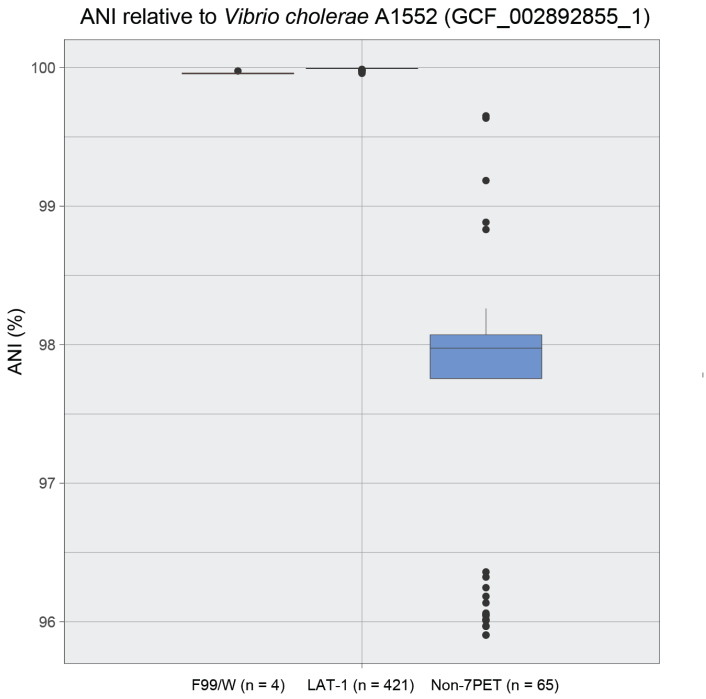


Figure 3.28 – ANI values for genomes sequenced in this study relative to the A1552 reference sequence. Samples were stratified by phylogenetic assignments.

3.5 – Discussion

A combination of factors mean that this study could only have been performed in Argentina. The limited introductions of 7PET sub-lineages into Latin America, the sole introduction of LAT-1 into Argentina, and the enhanced surveillance systems that were introduced in Argentina once cholera broke out in Peru all meant that the INEI culture collection was a comprehensive representation of the 1992-1998 Argentinian cholera epidemic in its entirety. Therefore, the number of isolates included in this project encompassed a very large proportion of the total cholera cases across the country and throughout the epidemic, presenting a unique opportunity to understand the dynamics of epidemic *V. cholerae* evolving over long periods, originating from point-source introductions. To the best of our knowledge, this is the largest genomic study to date that investigates a cholera epidemic in a single country. We also believe that it is the largest genomic analysis of any single bacterial pathogen in Argentina.

There are a number of observations to be made using these data, chief among which is the fact that it is clear that a single clone of *V. cholerae* O1, now known to be one sub-lineage of 7PET [189], was responsible for pandemic cholera in Argentina, in spite of the seasonal fluctuations and serotype variation observed (Figure 3.3) [274, 367, 399, 400]. These data also show that LAT-1 circulated amongst the countries at the Northern borders of Argentina during the early 1990s - for instance, cholera was first reported in Bolivia in August 1991 [401, 402], and Bolivian genomes from the early 1990s are mixed amongst the Argentinian genomes from the same period (Figures 3.14, 3.19). These genomic data also validate fundamental observations made by public health authorities during the cholera epidemics of the 1990s, such as that Argentinian cholera outbreaks were principally caused by *V. cholerae* O1 serotype Ogawa which had been shown by PFGE to be closely related to the Peruvian strain [208, 367, 403–405].

These data also explain historical observations on the variation of serotype during the 1990s epidemic across Latin America, and provide new insight into the microevolution of LAT-1 during the Argentinian cholera epidemic. *V. cholerae* Ogawa from Argentina in 1992 were shown to be closely related to Inaba isolates from Peru (Figure 3.19), and the shift in dominant serotype from Inaba to Ogawa that was observed in Peru and elsewhere in Latin America [208] represented variation within LAT-1, rather than a separate introduction of another strain (Figure 3.14A). Similarly, the outbreak of Inaba *V. cholerae* in Argentina in 1996 was also a

result of variation within LAT-1; the outbreak was caused by LAT-1 isolates in which *wbeT* had mutated from wild-type Ogawa genotype to an Inaba genotype (Q274trunc). This mutation is likely to have occurred in Argentina. The Q274trunc mutation is distinct from others identified within LAT-1, particularly the mutation associated with the Inaba phenotype in contemporaneous Mexican isolates (N165fs). These results strongly indicate that Argentinian cholera in 1996 was not caused by an introduction of a new Inaba (sub)lineage from elsewhere in Latin America; rather, LAT-1 *V. cholerae* that had already been introduced into Northern Argentina, or neighbouring countries, acquired a new Inaba genotype. In turn, the Argentinian cholera outbreak in 1997 was caused by a sub-clade of LAT-1 that was closely related to the 1996 Inaba clone. However, the topology of our phylogeny suggests that this was not a result of 'reversion' from the Inaba Q274trunc genotype to an Ogawa genotype (Figure 3.19). These data underline that Ogawa/Inaba phenotypic variation is not phylogenetically informative, and may be both misleading and inappropriate to use as an epidemiological marker.

Furthermore, the inclusion of non-7PET *V. cholerae* in our study has highlighted that a highly diverse population of the *V. cholerae* species existed in Argentina concurrently with the extremely invariant LAT-1 pandemic sub-lineage during the 1990s. It is suggested that these non-7PET bacteria, including serogroup O1 and non-O1 isolates, represent those *V. cholerae* that are truly endemic to Argentina, which are evolving locally, but lack the propensity or ability to cause global epidemics and to spread in the same way as 7PET. Therefore, it can be concluded that the reason that Latin America was cholera-free for 97 years was due solely to the absence of pandemic *V. cholerae* lineages from the continent. In the absence of comprehensive clinical data associated with these non-7PET isolates, it cannot be determined whether they are aetiological agents of cholera, or of a cholera-like illness. However, it is very clear that non-7PET *V. cholerae* were present in Argentina, and associated with disease at a low level, throughout the 1992-1998 cholera epidemic and thereafter (Figure 3.3).

In spite of the sustained circulation and dissemination of LAT-1 across Northern Argentina, an area of approximately 1.2 million km^2 (Figure 3.4), these data suggest that very little genetic change, in terms of SNVs, recombination, and gene gain/loss, occurred in this sub-lineage over a period of nearly six years. The invariance of LAT-1 is juxtaposed with the diversity observed in non-O1 *V. cholerae* in Argentina (Figures 3.14A, 3.16, 3.21, 3.24-3.28). As discussed in the Introduction to this chapter, although *V. cholerae* research has focused on studying epidemics and outbreaks, by definition, this tends to describe epidemic lineages. Non-7PET *V. cholerae*

are highly variable – within this dataset, as well as examples of local lineages of non-7PET *V. cholerae*, isolates were also identified which were confirmed microbiologically to be *Vibrio cholerae*, but were diverse phylogenetically (Figure 3.21) and as measured by ANI values (Figure 3.28). The pathology of the disease associated with these isolates – and whether virulence determinants such as T3SS contribute to this disease – is beyond the scope of this thesis but is the focus of future work, though evidence does suggests that T3SS do contribute to diarrhoea and disease caused by non-7PET bacteria [305]. With the caveat of a small sample size, it can also be observed that the clinical non-7PET isolates were enriched for the presence of T3SS (16/21 isolates).

It is particularly vital to understand the diversity of local, endemic *V. cholerae* that co-exist alongside 7PET during a cholera epidemic, because non-epidemic *V. cholerae* present in a country may contribute to disease that is symptomatic of cholera, but does not pose the same relative risk to public health as 7PET [189]. Similar observations have recently been made in China [396]. The relative risk of *V. cholerae* lineages should be accounted for in the magnitude of epidemic preparedness responses to such outbreaks. This is particularly relevant in the wake of the GTFCC commitment to reducing deaths from cholera by 90% before the year 2030 [9]. This campaign focuses on the control of cholera, the disease, rather than on controlling 7PET, the aetiological agent of pandemic cholera. As cholera control is implemented, countries experiencing a high incidence of cholera attributable to 7PET will see a decline in the number of cholera cases. Therefore, it is anticipated that as pandemic cholera reduces in magnitude, disease caused by non-7PET *V. cholerae* will become more visible, as was observed in Argentina and Latin America. By using genomics to differentiate pandemic and non-pandemic lineages for public health epidemic preparedness responses, concerted control efforts including epidemiologists, public health authorities and microbiology laboratories targeting 7PET specifically, and accounting for background levels of endemic non-7PET disease, could see epidemic cholera eliminated in Latin America.

To summarise, as well as describing the genomic history of cholera in Argentina during the 1990s, this chapter describes the stark contrasts can be observed during a country-wide cholera epidemic between the clonality of 7PET and the diversity of contemporaneously-isolated non-7PET *V. cholerae*. This suggests that to understand more about the disease, particularly the disease that can be caused by non-epidemic and non-pandemic *V. cholerae*, our efforts must

begin to turn from the exclusive study of pandemic *V. cholerae* towards a more holistic and concurrent analysis of pandemic and non-pandemic isolates.

This chapter has focused on analysing LAT-1 in detail, and has made high-level comparisons between the LAT-1 and species datasets. Clearly, although non-7PET *V. cholerae* continue to be associated with clinical cases of disease, they remain understudied. In Chapter 4, a detailed study of a small number of historically- and medically-important non-7PET non-O1 *V. cholerae* will be described, using both *in silico* and *in vitro* approaches. This will focus on characterising specific aspects of their genomes. Once these isolates have been described in fine-detail, subsequent chapters will be of broader scope, return to studying larger collections of genomes *in silico* with the aim of extrapolating the results of these analyses into the context of larger collections of diverse genomes. The non-7PET genomes described in this chapter will also be re-examined in subsequent chapters, to consider aspects such as antimicrobial resistance, virulence gene distribution, and plasmid replicons, in the context of additional diverse genomes and the data presented in Chapter 4.

Chapter 4

Long-read sequencing of modern and historical *V. cholerae*

Contribution statement

Nick Thomson supervised the work described in this chapter. The *V. cholerae* O139 genome sequence data were made available for this PhD by Firdausi Qadri. Cultures of NCTC strains were supplied by Sarah Alexander and Julie Russell. The pACYC184 plasmid stock used for molecular cloning was a gift from Francesca Short. Jake Turnbull assisted with the collation of NCTC internal records, and Mohammed-Abbas Fazal prepared gDNA from NCTC 30 batch 4 at PHE laboratories. Leanne Kane and I prepared samples for electron microscopy, which were imaged by Claire Cormie and David Goulding. Daryl Domman generated a Canu assembly for isolate 48853_G01.

I performed all experiments and genomic analyses, and produced all figures except for microscopy images.

Publication

Some figures and data used in this chapter have been published in the following articles:

Dorman MJ, Kane L, Domman D, Turnbull JD, Cormie C, Fazal M-A, Goulding DA, Russell JE, Alexander S & Thomson NR (2019). The history, genome and biology of NCTC 30: a non-pandemic *Vibrio cholerae* isolate from World War One. *Proceedings of the Royal Society B* **286** (1900): 20182025.

Dorman MJ & Thomson NR (2020). Community evolution: Laboratory strains in the era of genomics. *Microbiology* **166** (3): 233-238. Invited "insight review" article.

Dorman MJ*, Domman D*, Uddin MI*, Sharmin S, Afrad MH, Begum YA, Qadri F & Thomson NR (2019). High quality reference genomes for toxigenic and non-toxigenic *Vibrio cholerae* serogroup O139. *Scientific Reports* **9** (1): 5865. (* Joint first author)

4.1 – Overview

In Chapter 3, I used a collection of *V. cholerae* genomes from Argentina to perform a detailed analysis of the evolution of a sub-lineage of highly clonal and genomically-invariant pandemic *V. cholerae* O1. As has been mentioned previously (section 3.5), if research efforts are focused on 7PET alone, the diversity of the species will be neglected. To that end, Chapter 3 included a high-level characterisation of genomic differences between pandemic and non-pandemic *V. cholerae*. For instance, amongst just 61 non-O1 Argentinian *V. cholerae*, examples could be seen of isolates with ANI values suggesting that they were on the boundary of being classifiable as a new species (section 3.4.10; Figure 3.28).

The fact that the diversity of the *V. cholerae* species is juxtaposed to the clonality of 7PET indicates that to understand this species further, the sequencing and characterisation of non-pandemic *V. cholerae* is required. However, although thousands of *V. cholerae* have been sequenced to date, there are still very few high-quality and accurate genome assemblies for this species, particularly for non-pandemic non-O1 *V. cholerae* isolates. High-quality genome sequences are needed to perform robust comparative genomic studies, some of which have been discussed previously (Introduction, section 1.4.1). Such comparative studies are required to address one of the most fundamental questions in *V. cholerae* research – the need to describe what discriminates pandemic and non-pandemic *V. cholerae*.

Since exploring the differences between pandemic and non-pandemic *V. cholerae* was a major aim of this PhD (section 1.5), it was decided to use long-read sequencing to generate accurate assemblies for five important *V. cholerae* isolates. These comprise four recently-isolated *V. cholerae* of serogroup O139, and a fifth historical strain isolated in 1916. Long-read technologies were used in order to resolve problems that are specific to the sequencing of *V. cholerae*, such as the presence of repeats in the integron on chromosome 2 [49, 406], and the potential for multiple copies of the CTXφ prophage to be integrated into one or both chromosomes [54, 162]. Having access to long- and short-read data for the same sample also enable hybrid assembly approaches, which use both long- and short-reads obtained from the same gDNA preparation, and allow the benefits of both PacBio RSII and Illumina technologies to be combined to produce high-quality genome assemblies [316, 319, 407], to a minimum of 'improved high-quality draft' status [407].

An overview of the history of *V. cholerae* O139 has been presented in the Introduction (section 1.3.2.1). It is known that epidemic *V. cholerae* O139 form a sub-lineage of the 7PET pandemic lineage [158, 234]. Although a serogroup O139 isolate obtained from a patient in India during 1992, dubbed MO10 [241], has been sequenced and used for comparative genomics in the past [54, 133], a closed genome for this clinically- and epidemiologically-important 7PET sub-lineage had not been reported prior to this PhD work. In this project, PacBio sequences for three toxigenic and one non-toxigenic *V. cholerae* O139 were assembled and analysed [244]. All four isolates were of clinical origin [244].

V. cholerae is a pathogen controlled under Schedule 5 of the Anti-Terrorism, Crime and Security Act (ATCSA) in the United Kingdom [408]. This means this bacterium must be worked on in a laboratory that meets specific security requirements. The only facilities at WSI that meet these requirements are a subset of our Containment Level 3 (CL3) laboratories. Therefore, as part of establishing a CL3 laboratory at WSI for conducting research into *V. cholerae*, a reliable protocol for the efficient extraction of high molecular weight gDNA from this species was required (Methods, section 2.2.5). Large fragments of gDNA that have not been mechanically sheared are required for long-read sequencing. The first *V. cholerae* isolate chosen to be cultured for PacBio sequencing using this methodology was NCTC 30, the oldest *V. cholerae* accessioned by NCTC, and to our knowledge, the oldest live isolate of this species that is publicly available for research. NCTC 30 is a unique scientific curiosity, as well as an important historical *V. cholerae* isolate.

4.2 – Specific aims

The work described in this chapter aimed to

1) Assemble, annotate, and characterise *V. cholerae* O139 genomes sequenced with long-read technologies, to generate reference assemblies for this medically-important sub-lineage of 7PET,
2) Sequence to completion the genome of NCTC 30, an historical curiosity and an important example of a non-O1/O139 *V. cholerae* which caused 'choleraic diarrhoea', and
3) Capitalise on having access to a live culture of NCTC 30 to validate genomic observations experimentally, and to optimise experimental protocols for use at CL3.

4.3 – Results

4.3.1 – Closed genome assemblies for toxigenic and non-toxigenic V. cholerae *O139*

Four previously-reported *V. cholerae* of serogroup O139 [244] had been re-sequenced using PacBio RSII technology prior to the beginning of this project, in order to generate reference-grade assemblies for genomic analyses. Illumina short-reads were also available for these isolates [244]. *De novo* assemblies of these PacBio reads were generated and annotated as described in Chapter 2.

Using the long-reads for each of these genomes, single contig assemblies were produced for each of the two chromosomes in each isolate, which were then corrected and circularised (Methods, section 2.1.2.2). The short-read data for each isolate were used to correct the assemblies, but this correction step did not make any improvements to the PacBio-only assemblies. Accordingly, it was decided to proceed with using these long-read assemblies for subsequent analyses in this chapter. Coverage percentages and other summary statistics for these assemblies are listed in Table 4.1.

Internal sequence ID	**Sample name**	**CTXφ present?**	**Genome size (bp)**	**Coverage of *de novo* assembly with long reads (%)**	**Coverage of N16961 (%)**	**SNVs relative to N16961**
48853_F01	MP_070116	No	4123525	165.76	58.5	122865
48853_G01	P_0684000	Yes	4092641	170.56	97	271
48853_H01	ICVB_2236_02	Yes	4092645	147.29	97	270
48853_A02	SMIC_67_01	Yes	4092644	165.65	97	274

Table 4.1 - Summary statistics for four closed *V. cholerae* O139 assemblies. Potentially recombined SNVs were not excluded from this analysis because 48853_F01 was extremely genetically distant from N16961. Modified from [409]. The HGAP assembler assembled the reads from sample 48853_G01 into three contigs. Re-assembling this sample with Canu v1.1 [410] produced a two-contig assembly, which was used for subsequent analysis.

4.3.2 – Phylogenetic position of toxigenic V. cholerae *O139*

The phylogenetic position of the three toxigenic sequences was confirmed by placing these isolates into context with additional *V. cholerae* O139 sequences [242, 244] and other 7PET

genomes [234]. A maximum-likelihood phylogeny was calculated from an alignment of 1,630 non-recombinant SNVs, which had been identified by mapping short-read data to the N16961 reference sequence (Figure 4.1). Since the three toxigenic isolates varied in length by four bases at most (Table 4.1), and were phylogenetically identical to one another, just one of the assemblies (48853_H01) was chosen to be used as an exemplar sequence for all subsequent comparative analyses.

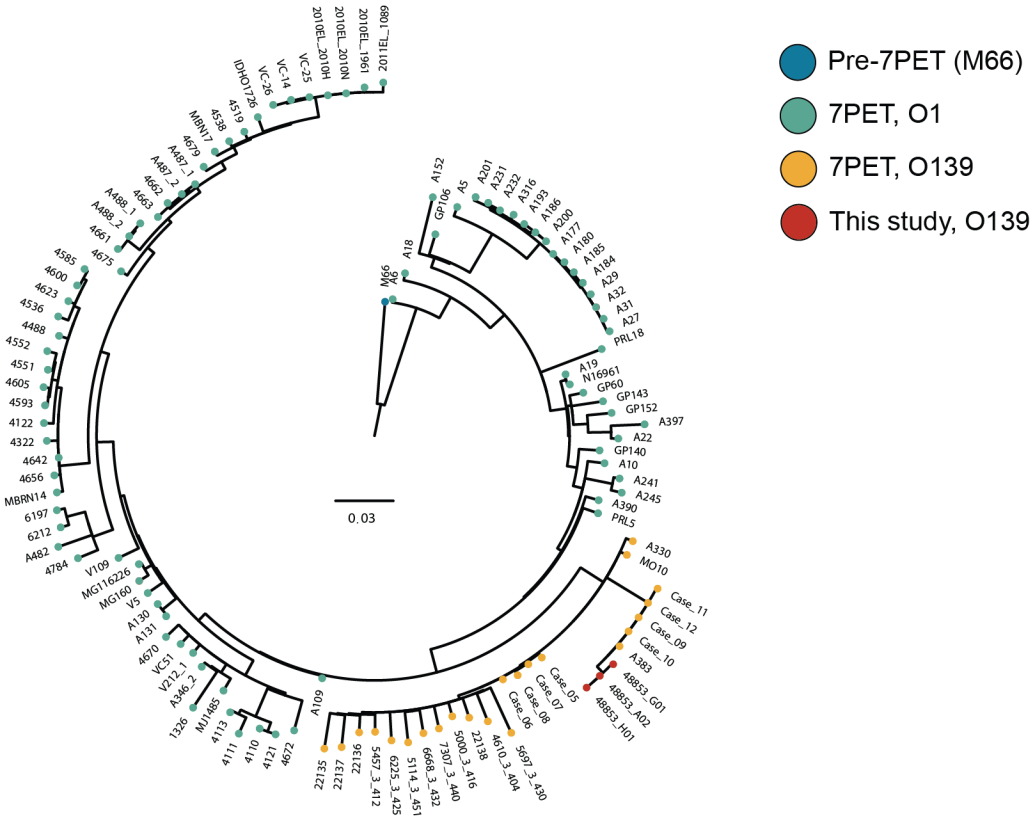


Figure 4.1 – Maximum-likelihood phylogenetic tree of 7PET. Phylogeny rooted on the M66 outgroup sequence. Scale bar denotes the number of substitutions *per* variable site. This phylogeny was computed under the GTR+Gamma model using RAxML v8.2.8 with 500 bootstrap replicates [411]. Modified from [409].

4.3.3 –CTXφ prophage sequences in toxigenic V. cholerae *O139*

Previous work had reported that multiple types and arrangements of CTXφ prophage had been detectable in *V. cholerae* O139 chromosomes, depending on the year and location from which an isolate was isolated [230]. In particular, isolates obtained from Calcultta in 1996 had been reported to harbour two types of prophage, CTXφ^{cal} and CTXφ$^{El\ Tor}$ in the configuration CTXφ$^{El\ Tor}$ - CTXφ^{cal} - CTXφ^{cal} [412–414]. However, the assembly for MO10, used by many as an O139 reference isolate [241] was insufficiently well-assembled to be confident of the

context of CTXφ in this laboratory strain; this assembly is currently available in 84 contigs (accession # GCA_000152425.1). Similarly, the short-read data from the three more recently sequenced *V. cholerae* O139 isolates could not be assembled across the CTXφ region into a single contig, and in those assemblies, only one of the two *ctxB* genes was identifiable (the Illumina assemblies for 48853_G01 and 48853_H01 contained only a *ctxB4* allele in a small contig, and 48853_A02 contained *ctxB5* in a larger contig).

Using the long-read assemblies generated here, it was evident that 48853_H01 contains three CTXφ prophages arranged in tandem, at the same integration site on the larger chromosome of N16961, between the *VC_1450* and *VC_1467* loci (Figure 4.2). This was the case in all three toxigenic *V. cholerae* O139 sequenced here (data not shown). In order to verify the presence of multiple copies of CTXφ in these genomes, the Illumina reads previously reported for this isolate were mapped to the N16961 reference and the coverage of these mapping data over the CTXφ region were visualised (Figure 4.3). The CTXφ region was covered 2-3 times as highly as the surrounding chromosome, supporting further the existence of multiple CTXφ copies in these *V. cholerae* O139 genomes.

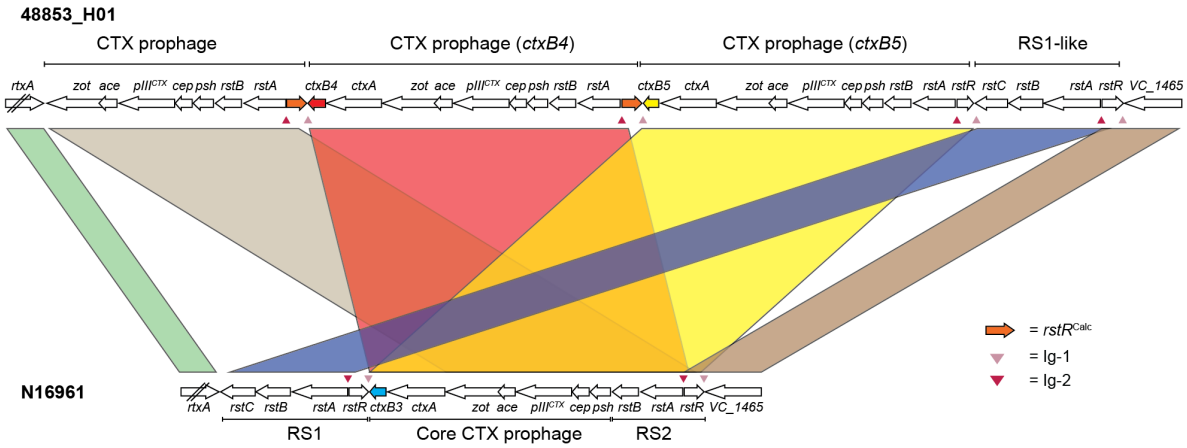


Figure 4.2 – Comparison of the CTXφ region in assembly 48853_H01 and N16961. Region annotated as *per* [99]. Reproduced from [409].

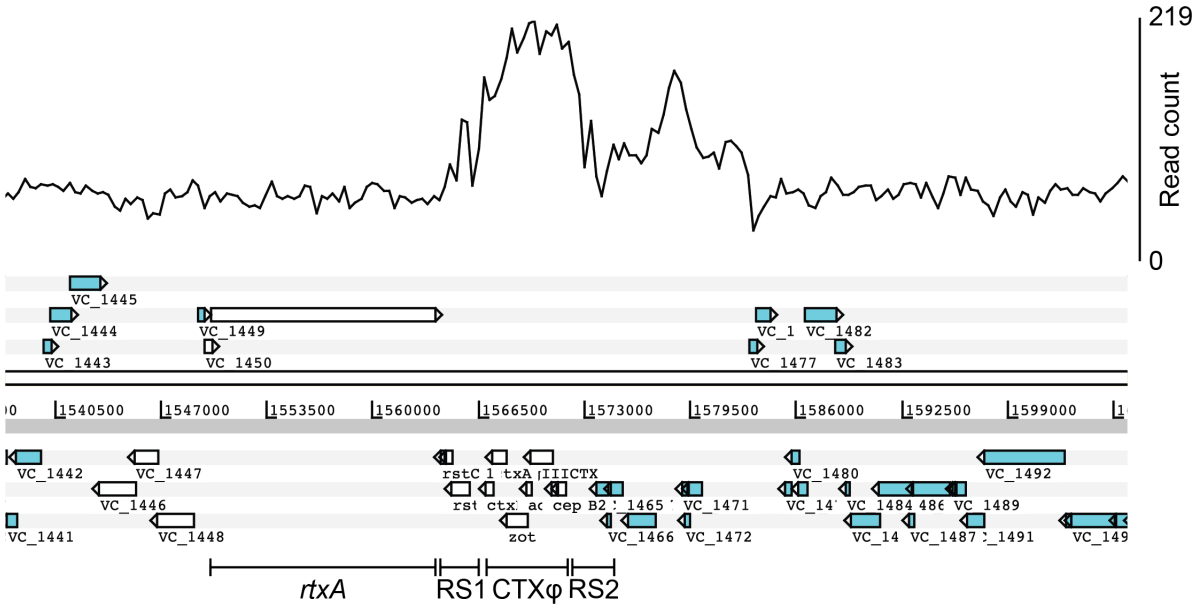


Figure 4.3 – Validating the presence of multiple CTXφ copies by mapping. Figure depicts the depth of short-reads for *V. cholerae* O139 isolate 48853_H01 mapped to the N16961 reference sequence. Mapping coverage was plotted using Bamview and Artemis [339, 356]. Annotation as described in Figure 4.2. Reproduced from [409].

Two of these repeats harboured distinct *ctxB* alleles, *ctxB4* and *ctxB5*. The *ctxB* gene closest to *rtxA* in these assemblies was a *ctxB5* allele, and the second *ctxB* was a *ctxB4* allele. Both *ctxB* alleles have been found in *V. cholerae* O139 strains previously [415, 416]. The third CTXφ repeat was partial, and lacked the *ctxAB* operon while comprising the genes between and including *zot* and *rstA*, and an *rstR* open reading frame corresponding to $rstR^{Calc}$ [417] (Figure 4.2). A complete *attL* sequence was identified adjacent to the *VC_1465* locus in 48853_H01 [96]. The phage sequence in the *attR* site adjacent to *rtxA* is not identical to that reported by Huber and Waldor [96], although the *attR* element does contain the central recombination motif and the residual bacterial *attB* sequence (see section 1.2.3 for an overview of CTXφ integration mechanisms). Although it has been reported that *V. cholerae* O139 can harbour more than one type of CTXφ phage simultaneously [230, 413, 414, 417], we were unaware of previous reports of multiple *ctxB* alleles co-existing in the same *V. cholerae* genome.

The *ctxB4* and *ctxB5* alleles differ from one another at two nucleotide positions; 83 and 115 (Figure 4.4A). These produce corresponding non-synonymous substitutions at amino acids 28 and 39 (Figure 4.4B). Additional confirmation that these alleles co-exist in the *V. cholerae* O139 genomes was obtained by inspecting manually the sequences of reads mapped to this

locus (Figure 4.5), and from the fact that the assemblies for all three toxigenic *V. cholerae* O139 harboured both *ctxB* alleles, in CTXφ that were in the same chromosomal arrangement in all three genomes (Figure 4.2). It was also confirmed by manual inspection of the assemblies that no CTXφ prophage were present in the smaller chromosome in these isolates.

A

```
                1
ctxB1       ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC
ctxB3/B3b   ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC
ctxB4       ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC
ctxB5       ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC

               71
ctxB1       AAAATATTAC TGATTTGTGT GCAGAATACC ACAACACACA AATACATACG CTAAATGATA AGATATTTTC
ctxB3/B3b   AAAATATTAC TGATTTGTGT GCAGAATACC ACAACACACA AATATATACG CTAAATGATA AGATATTTTC
ctxB4       AAAATATTAC TGATTTGTGT GCAGAATACC ACAACACACA AATATATACG CTAAATGATA AGATATTTTC
ctxB5       AAAATATTAC TGCTTTGTGT GCAGAATACC ACAACACACA AATACATACG CTAAATGATA AGATATTTTC

              141
ctxB1       GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AACTTTTCAA
ctxB3/B3b   GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AATTTTTCAA
ctxB4       GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AACTTTTCAA
ctxB5       GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AACTTTTCAA

              211
ctxB1       GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA
ctxB3/B3b   GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA
ctxB4       GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA
ctxB5       GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA

              281
ctxB1       GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT
ctxB3/B3b   GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT
ctxB4       GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT
ctxB5       GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT

              351
ctxB1       TGCCGCAATT AGTATGGCAA ATTAA
ctxB3/B3b   TGCCGCAATT AGTATGGCAA ATTAA
ctxB4       TGCCGCAATT AGTATGGCAA ATTAA
ctxB5       TGCCGCAATT AGTATGGCAA ATTAA
```

B

Signal sequence (21 AA)

```
                1
CtxB1       MIKLKFGVFF TVLLSSAYAH GTPQNITDLC AEYHNTQIHT LNDKIFSYTE SLAGKREMAI ITFKNGATFQ
CtxB3/B3b   MIKLKFGVFF TVLLSSAYAH GTPQNITDLC AEYHNTQIYT LNDKIFSYTE SLAGKREMAI ITFKNGAIFQ
CtxB4       MIKLKFGVFF TVLLSSAYAH GTPQNITDLC AEYHNTQIYT LNDKIFSYTE SLAGKREMAI ITFKNGATFQ
CtxB5       MIKLKFGVFF TVLLSSAYAH GTPQNITALC AEYHNTQIHT LNDKIFSYTE SLAGKREMAI ITFKNGATFQ

               71
CtxB1       VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
CtxB3/B3b   VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
CtxB4       VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
CtxB5       VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
```

Figure 4.4 – Alignment of *ctxB* and CtxB variants. Highlighted *ctxB* allelic variations (A) give rise to non-synonymous variations in encoded CtxB proteins (B). The *ctxB1* allele is associated with Classical *V. cholerae*, and *ctxB3* with wave 1 7PET *V. cholerae* such as N16961 [59, 97, 234]. None of these mutations are located in the region of *ctxB* which encodes the N-terminal signal sequence (annotation taken from [77]).

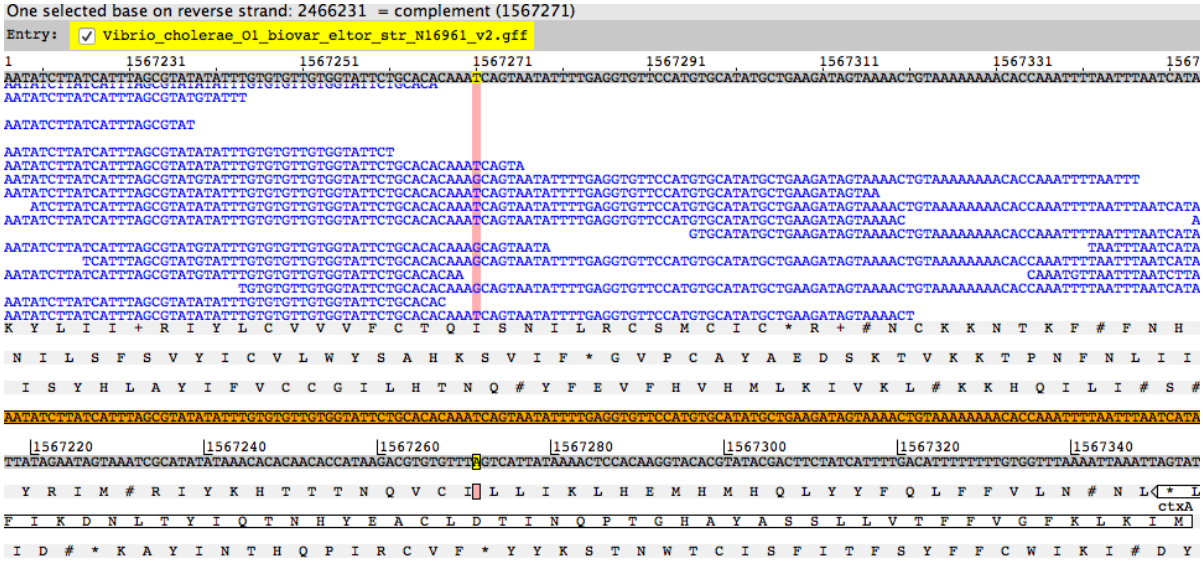


Figure 4.5 – Reads corresponding to both *ctxB4* and *ctxB5* from toxigenic *V. cholerae* O139 map to N16961. In this example, the base at position 83 in *ctxB* has been highlighted (see Figure 4.4). These reads are mapped to the opposite strand to the annotated *ctxB*.

Since multiple CTXφ configurations have been reported in *V. cholerae* O139, and in order to broaden our observations, a set of 50 *V. cholerae* O139 genomes that were obtained during a study of intra-household cholera transmission in Dhaka, Bangladesh were re-examined [235]. These formed two sub-lineages within 7PET (dubbed B3 and B4) and were isolated in 2002 and 2003, pre-dating the isolates sequenced in this study. Forty-one of these isolates harboured both *ctxB4* and *ctxB5* alleles simultaneously, all of which were in lineage B4. The short-read data for the genomes sequenced in this chapter had been included in the Dhaka household transmission study, and were closely related to B4 [235]. The remaining 9 only harboured one *ctxB* allele, *ctxB4*. All of these were in lineage B3. This suggests that there is genomic complexity and population structure within the O139 sub-lineage of 7PET that has been previously-hidden due to the lack of an accurate reference sequence for the sub-lineage.

4.3.4 – Genomic island complements of V. cholerae *O139 isolates*

Having seen variation in CTXφ, the complement of other genomic islands contained in the three toxigenic and one non-toxigenic *V. cholerae* O139 isolates' genomes was determined. Examined were the canonical pathogenicity islands, prophages, the SXT ICE*Vch*Ind4 element, and other genomic islands that have been reported previously in *V. cholerae* [54, 55, 133, 134, 141]. VPI-1 and VSP-2 were present in all three toxigenic *V. cholerae* O139 isolates (Table

4.2; Figure 4.6). Moreover, VPI-2 was severely truncated to the point of absence in these three isolates. This truncation has been described previously in MO10 and is a hallmark of toxigenic *V. cholerae* O139 [54, 418] (Table 4.2; Figure 4.6). VPI-1 was partially present in the genome of the non-toxigenic isolate (Table 4.2).

Sample Name	**VSP-1 (*VC_0174-VC_0186*)**	**VSP-2 (*VC_0489-VC_0517*)**	**VPI-1 (*VC_0809-VC_0848*)**	**VPI-2 (*VC_1757-VC_1810*)**	**CTXφ (*VC_1451-VC_1465*)**	**SXT (*VC_0659* insertion)**
48853_F01	Absent	Absent	Partially present (deletion of *VC_0817-VC_0848*)	Absent	Absent	64% match to ICE*Vch*Ind4
48853_G01	Present, and duplication on chr2	Present	Present	Deletion of *VC_1761-1787*	Present, in more than one copy	100% match to ICE*Vch*Ind4
48853_H01	Present, and duplication on chr2	Present	Present	Deletion of *VC_1761-1787*	Present, in more than one copy	100% match to ICE*Vch*Ind4
48853_A02	Present, and duplication on chr2	Present	Present	Deletion of *VC_1761-1787*	Present, in more than one copy	100% match to ICE*Vch*Ind4

Table 4.2 – Presence and absence of select genomic islands in *V. cholerae* O139 genome assemblies. Similarity percentages were obtained by comparing SXT element sequences to that of ICE*Vch*Ind4 using BLASTn. chr2 = chromosome 2. Modified from [409].

A genomic island integrated into the *VC_0659* locus (encoding peptide chain release factor 3) was detected in each of the three toxigenic *V. cholerae* O139 assemblies. This island was identical to SXT (also known as ICE*Vch*Ind4 [156]) and was integrated into the same locus as is SXT in MO10 [156]. An insertion into *VC_0659* was also identified in the non-toxigenic O139 genome assembly, which was 64% identical to ICE*Vch*Ind4 (Table 4.2). These observations are fully consistent with previous data on genomic island distribution in the MO10 genome [54].

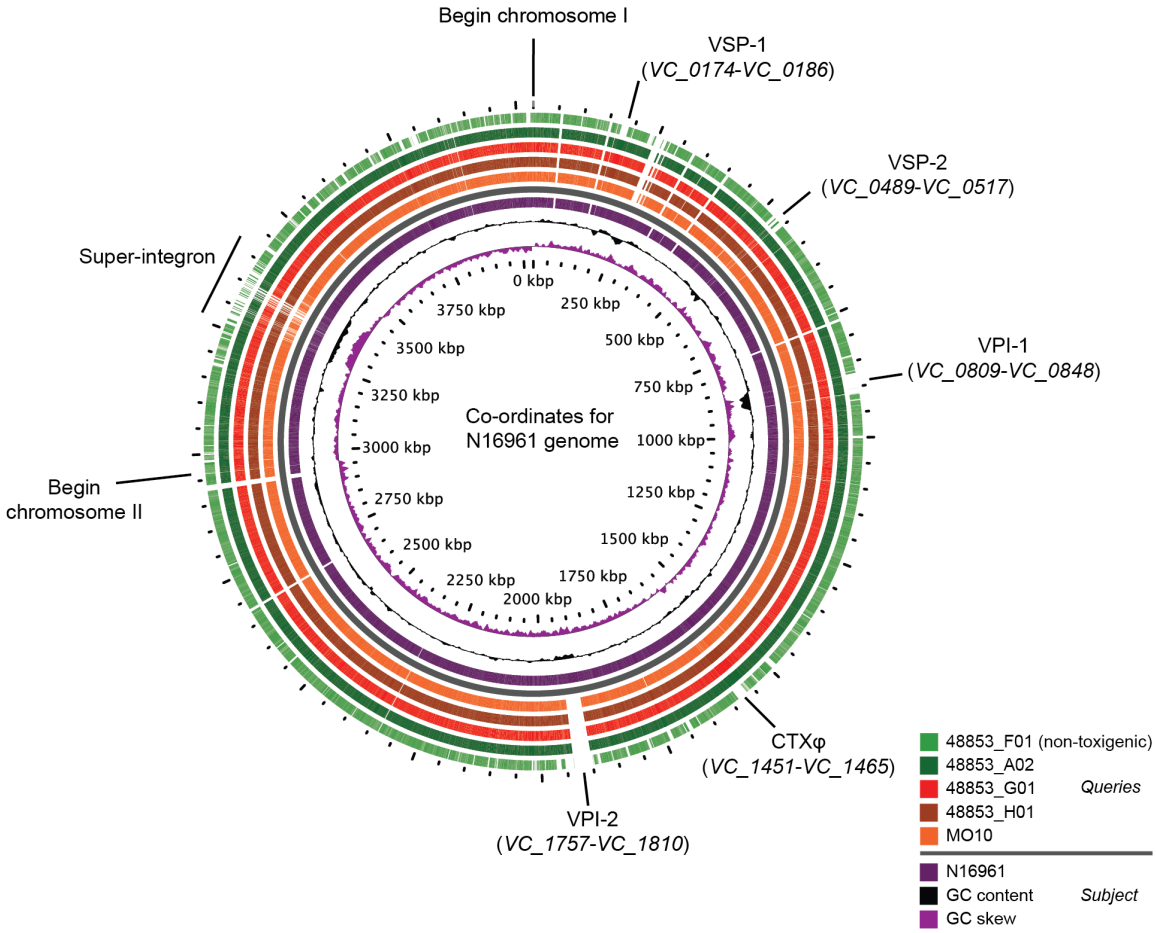


Figure 4.6 – BLAST atlas illustrating the location, presence, and absence of key genomic islands in *V. cholerae* O139 assemblies relative to the N16961 reference sequence. The sequences of both N16961 chromosomes were concatenated to produce this figure. Reproduced from [409].

VSP-1 was integrated on the larger chromosome between genes *VC_0173* and *VC_0187* in the three toxigenic *V. cholerae* O139 isolates, as it is in N16961 [54] (Figure 4.6; Table 4.2). However, a 14.3 kb sequence of DNA was also detected on the smaller chromosome of each of the toxigenic isolates, integrated between *VC_A0695* and *VC_A0696*, that was 99% identical at the nucleotide level to VSP-1 (*VC_0175* to *VC_0186*; Figure 4.7A). This strongly suggested that a second copy of the VSP-1 element was present on the second chromosome in each of these genomes. In order to verify this, the short-reads for each isolate were mapped to the N16961 reference genome the read depth was plotted for VSP-1 relative to the surrounding genome (Figure 4.7B).

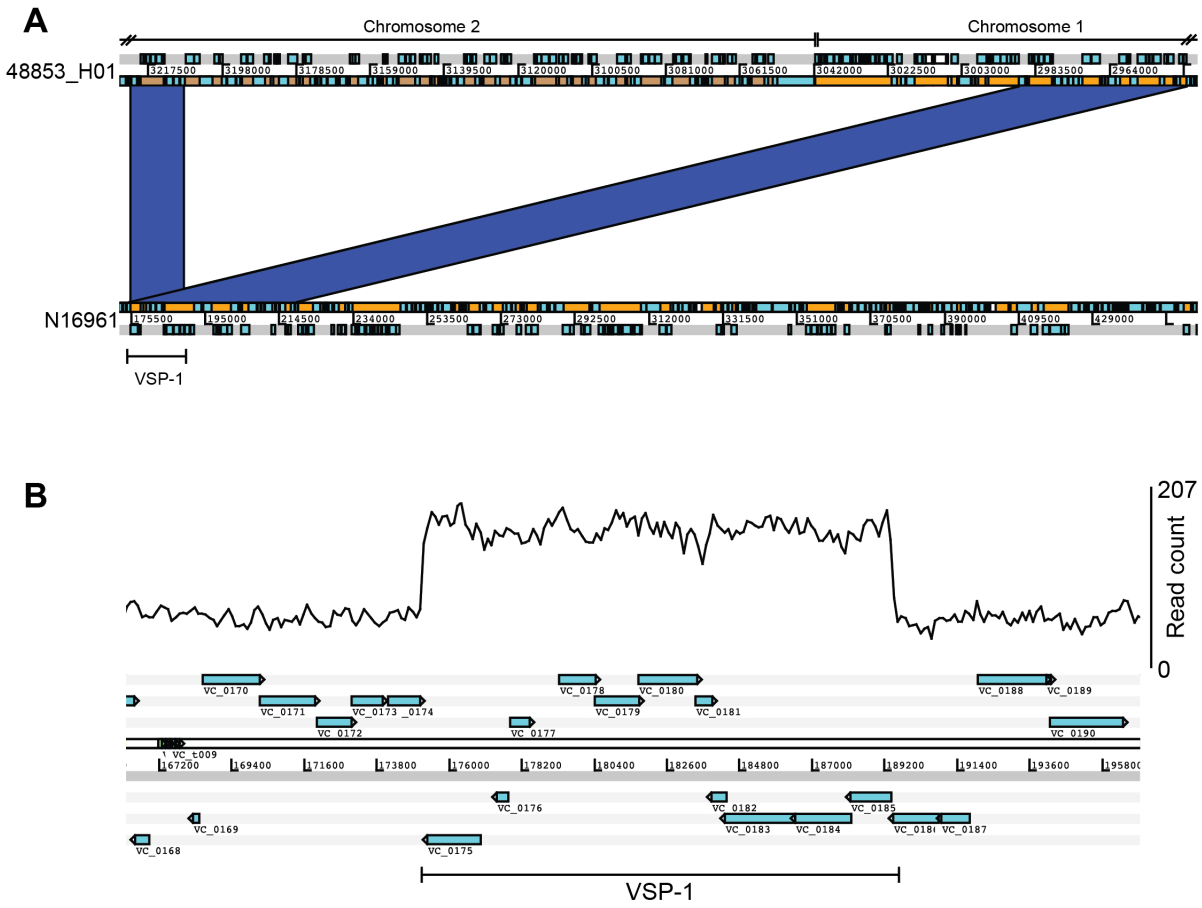


Figure 4.7 – Presence of VSP-1 on both chromosomes of *V. cholerae* O139. DNA that was homologous to VSP-1 was detected on both chromosomes of 48853_H01 (A). Mapping short-reads from 48853_H01 to N16961 confirmed that there were approximately double the number of sequencing reads for VSP-1 relative to the surrounding chromosome (B). Modified from [409].

This VSP-1 duplication was detected in each of the three toxigenic *V. cholerae* O139 genome assemblies (Table 4.2). It is known that VSP-1 is capable of excising from the larger *V. cholerae* chromosome [141], and it has been previously reported that the *V. cholerae* O1 isolate MJ-1236 (the Matlab variant) harbours a second copy of VSP-1 integrated between *VC_A0695* and *VC_A0696* [419]. Separately, Grim *et al*. identified a single clinical isolate of *V. cholerae* O139 from Bangladesh that appeared to harbour an insertion between *VC_A0695* and *VC_A0696* that resembled VSP-1, but this isolate was not described further [419]. This is likely to be the same phenomenon observed and confirmed to be present in these three *V. cholerae* O139 genomes.

4.3.5 – Antimicrobial resistance and accessory virulence determinants

Since the ambition of this project was to use these assemblies as reference sequences for future studies of *V. cholerae* O139, having made comparisons to N16961, these assemblies were then compared to the MO10 sequence. This allows for consistency amongst these data and comparisons with previous publications (Figure 4.8). MO10 harbours the VSK prophage GI-16 [54], which N16961 lacks. GI-16 is also absent from the four sequences in this study (Figure 4.8). Likewise, MO10 harbours a kappa prophage (genomic island GI-11 [54]), which is neither present in N16961 nor in the O139 sequences characterised here (Figure 4.8). Importantly, MO10 does not appear to harbour the second VSP-1 copy on chromosome 2 identified in these three isolates, which were isolated more recently than MO10.

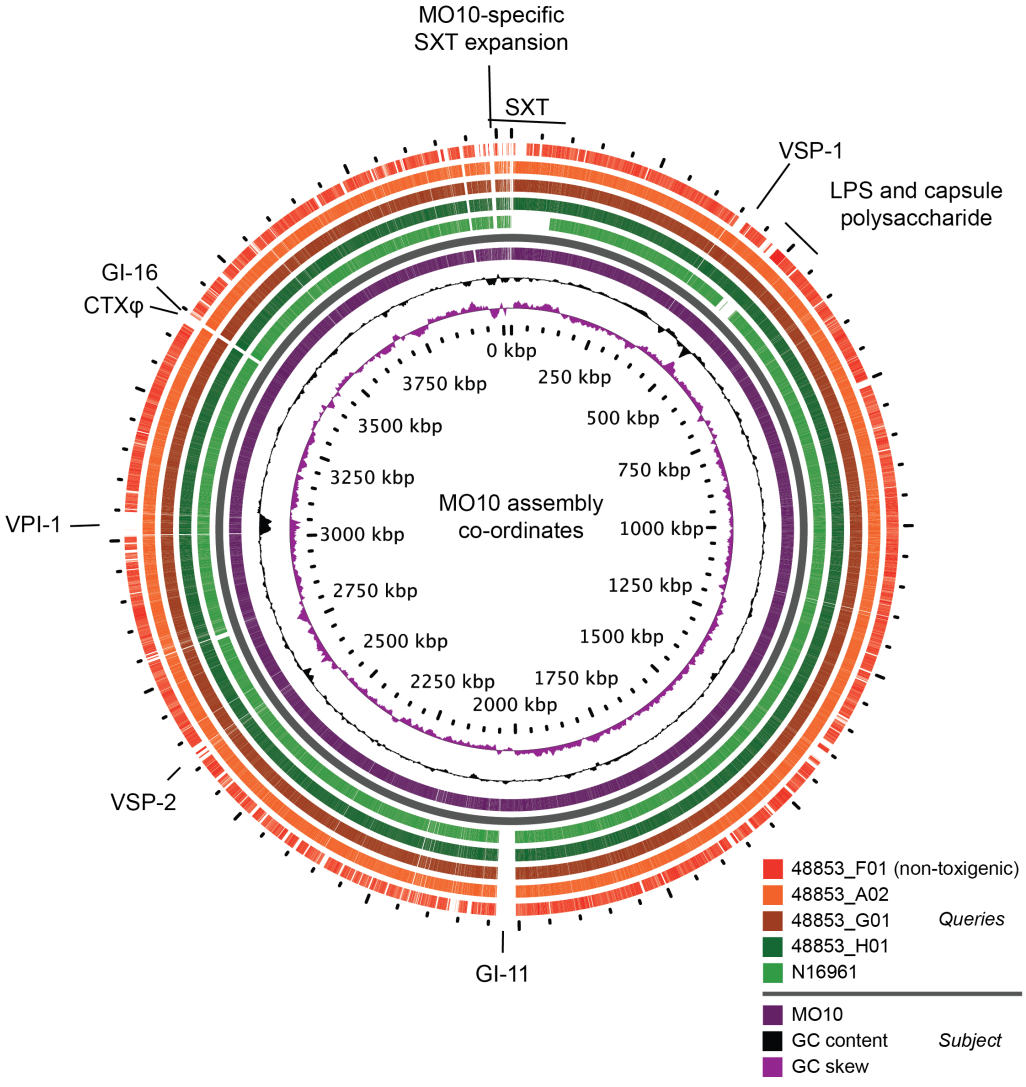


Figure 4.8 – BLAST atlas comparing *V. cholerae* O139 assemblies to MO10. The absence of O139-specific LPS and capsule genes from N16961 is evident, as is the absence of pathogenicity islands and CTXφ from the non-toxigenic isolate. All MO10 contigs were concatenated to produce this figure. Reproduced from [409].

MO10 harbours an SXT element [156] which is expanded relative to that found in these genomes (Figure 4.8). The MO10-specific SXT expansion includes genes conferring resistance to the antimicrobials streptomycin (*strAB*), sulfamethoxazole (*sul2*), trimethoprim (*dfr18*), and chloramphenicol (*floR*). Accordingly, the assemblies for all four *V. cholerae* O139 genomes were scanned for antimicrobial resistance genes. No antimicrobial resistance genes were detected in the non-toxigenic 48853_F01 genome. The three toxigenic isolate genomes also do not contain antimicrobial resistance genes with the exception of one *catB9* gene that is common to almost all 7PET *V. cholerae* [158] (Chapter 3, Figure 3.18), and is known not to render isolates resistant to phenicols [420]. These results agree with the original antimicrobial sensitivity testing of these isolates, which found that they were resistant only to nalidixic acid [244]. Examination of the *gyrA* and *parC* genes in these sequences confirmed that these four isolates harbour mutations predicted to result in an S83I mutation in GyrA. The non-toxigenic isolate 48853_F01 contains an additional mutation in *gyrA* (predicted to cause substitution A171S) and a S85L mutation in *parC* (predicted to cause an S85L substitution in ParC). All of these mutations are associated with nalidixic acid resistance in *V. cholerae* [155].

The four genome assemblies were also scanned for the presence of *V. cholerae* accessory virulence genes, with the specific objective of determining whether candidate virulence genes were present in the genome of the otherwise non-toxigenic *V. cholerae* O139 isolate, given that this isolate was obtained from a patient suffering from severe diarrhoea [244]. However, no known virulence determinants were detected in the 48853_F01 genome assembly other than those typically found across the *V. cholerae* species [1] (Table 4.3).

Accessory virulence gene	Present in 48853_F01	Present in 48853_G01	Present in 48853_H01	Present in 48853_A02
ToxR (*VC_0984*)	Yes	Yes	Yes	Yes
Zona occludens toxin, Zot (*VC_1458*)	No	Yes	Yes	Yes
Accessory cholera enterotoxin, Ace (*VC_1459*)	No	Yes	Yes	Yes
Haemolysin, *hlyA* (*VC_A0219*)	Yes	Yes	Yes	Yes
Mannose-sensitive haemagglutinin, MSHA (*VC_0398..VC_0414*)	Yes	Yes	Yes	Yes
MARTX toxin, *rtxA* (*VC_1451*)	Yes	Yes	Yes	Yes
MARTX toxin accessory gene, *rtxC* (*VC_1450*)	Yes	Yes	Yes	Yes
HA/protease, *hapA* (*VC_A0865*)	Yes	Yes	Yes	Yes
Heat-stable enterotoxin NAG-ST (Accession # M85198.1)	No	No	No	No
Type III secretion system from *V. cholerae* AM_19226 (typically present *in lieu* of VPI-2; accession # AATY01000000)	No	No	No	No

Table 4.3 – Accessory virulence genes in *V. cholerae* O139. Locus IDs from the N16961 reference, or accession numbers for sequences absent from N16961, are provided. Reproduced from [409].

4.3.6 – Phylogenetic position of non-toxigenic V. cholerae *O139*

It had been reported previously that isolate 48853_F01 was phylogenetically distantly related to the three toxigenic isolates [244]. In order to contextualise this isolate, the genomes of 61 additional *V. cholerae*, and those of three non-cholera Vibrios, were used to calculate an alignment of 2,103 core genes. From this sequence alignment, 168,476 SNVs were identified, which were used to calculate a core-gene phylogeny for these isolates (Figure 4.9).

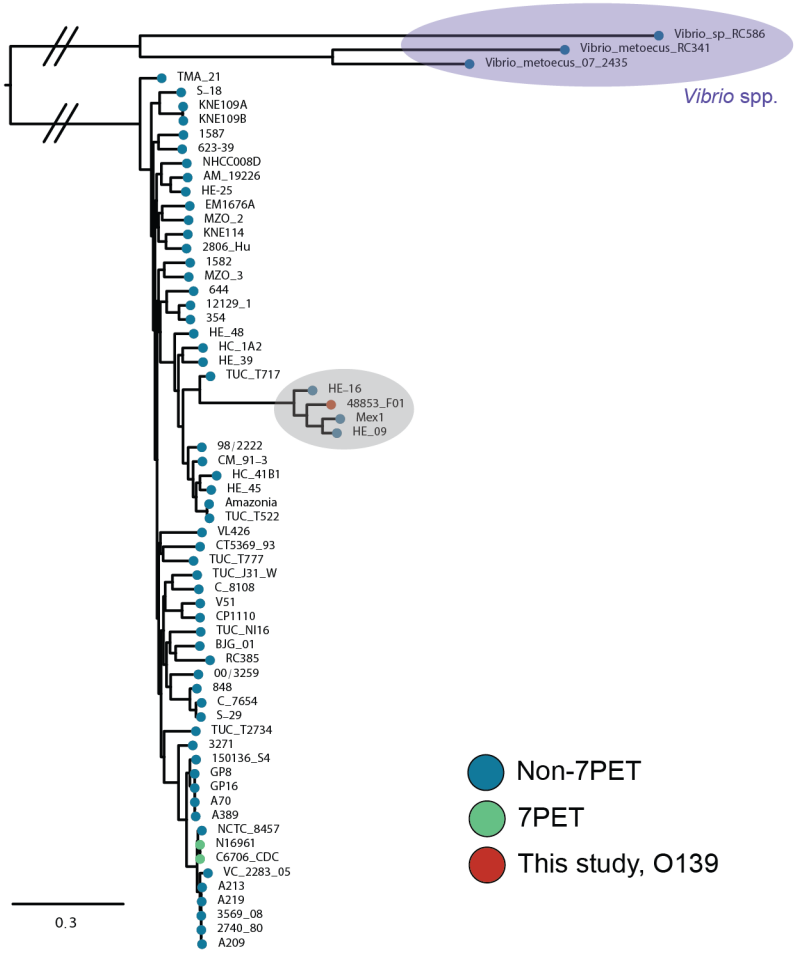


Figure 4.9 – A phylogeny of non-7PET *V. cholerae*. Rooted on three *Vibrio* spp. genomes. Hatch marks denote branches which have been manually shortened for illustrative purposes. This phylogeny was computed under the GTR+Gamma model using RAxML v8.2.8 with 500 bootstrap replicates [411]. Modified from [409].

The phylogenetic analysis showed that 48853_F01 forms a cluster with three isolates that are distantly related to the three toxigenic *V. cholerae* O139, falling outside of the 7PET lineage to which they belong (Figures 4.1, 4.9; grey disc). This cluster includes two Haitian non-O1 *V. cholerae* isolates from 2010 and a Mexican isolate from 1991 – unfortunately, serotype data were not recorded for these isolates (Figure 4.9). Moreover, several Argentinian *V. cholerae* sequenced in Chapter 3 were found to cluster with these sequences (shown in Figure 3.21).

Given the phylogenetic position of this non-toxigenic isolate, the capsule and lipopolysaccharide (LPS) biosynthesis loci in this genome were compared to the MO10 sequence and to those *V. cholerae* O139 that belonged to the 7PET lineage. Using closed genome assemblies generated here, it was confirmed that the three toxigenic strains contain O139 LPS operons that strongly resemble that found in MO10 (Figure 4.10). The equivalent

region in the non-toxigenic isolate 48853_F01 is less similar, although this strain exhibits a strong O139-positive phenotype using the rapid dipstick assay and slide agglutination tests [244]. It was also noted that the three non-O1 Haitian and Mexican *V. cholerae* which cluster with the non-toxigenic O139 isolate share capsule biosynthesis genes with 48853_F01, but they do not harbour the same LPS operon (Figures 4.9, 4.10). These isolates therefore are unlikely to be *V. cholerae* of serogroup O139.

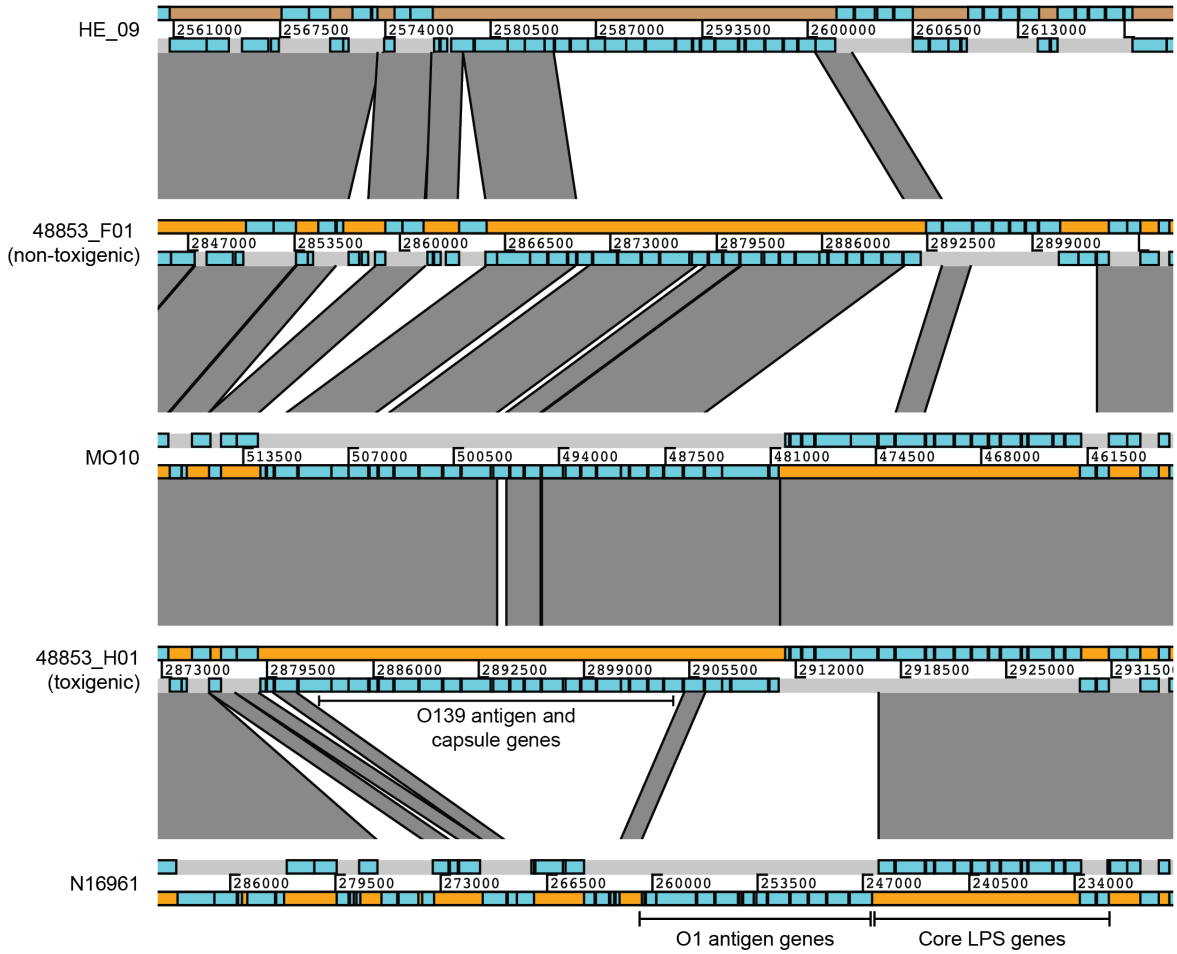


Figure 4.10 – O-antigen chromosomal loci in select *V. cholerae*. The loci known to encode the O1 and O139 antigens, as well as capsule biosynthesis genes, are indicated. HE_09 is phylogenetically related to the non-toxigenic *V. cholerae* O139 (Figure 4.9) but lacks the O139 biosynthesis operon. Reproduced from [409].

Serendipitously, the phylogenetic position of the non-toxigenic isolate 48853_F01 was made even more intriguing in the context of work on NCTC 30, the genomic and experimental characterisation of which is now presented.

4.3.7 – NCTC 30 genome sequencing and assembly

NCTC 30 was isolated in 1916 and was accessioned into NCTC in 1920 (Figure 4.11). This *V. cholerae* was originally named "Martin 1", and was isolated from a British soldier convalescent in Egypt against the background of World War 1 (WW1). This was a period in history during which cholera's epidemic potential was both recognised and feared [421, 422]. In spite of this, the British Expeditionary Forces remained largely free of cholera throughout this period. It has been estimated that the British Expeditionary Forces incurred 11,096,338 casualties during WW1, but reportedly experienced just 1,918 cholera cases in 1916, 209 cases in 1917, and 450 cases in 1918 [423]. Of these cholera patients, 49 died in 1917, and 106 died in 1918 [423].

NCTC 30 has been described as being 'probably' serogroup O2, and its biochemical profile has been described as part of its maintenance as the thirtieth strain to be accessioned into the NCTC culture collection (Figure 4.11). NCTC 30 stocks have been maintained since the strain was accessioned into the collection (Figure 4.12).

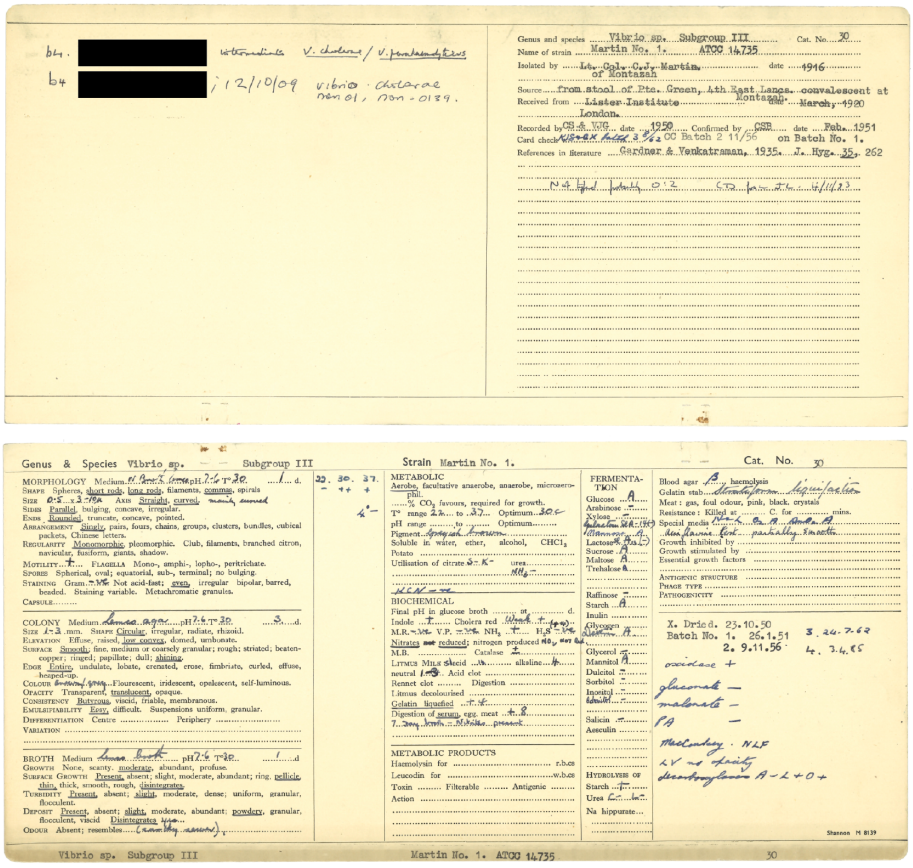


Figure 4.11 – The NCTC 30 check card. The card details the provenance of the isolate as well as aspects of its original biochemical and microbiological characterisation. Reproduced from supplementary material of [424].

For this PhD, a freeze-dried culture of NCTC 30 was revived and sequenced to completion using both long- and short-read technologies. NCTC has produced four lyophilised batches of NCTC 30 to date (Figure 4.11, 4.12), and the culture sequenced in this study was a derived from the earliest batch of lyophilised bacteria available for research (batch 3, lyophilised in 1962; Figures 4.11; 4.12). This was to minimise the risk of sequencing a culture that had acquired mutations as a consequence of long-term passage or maintenance. Molecular and genetic experiments were then carried out, to characterise further the biology of this historical isolate, and to validate genomic results.

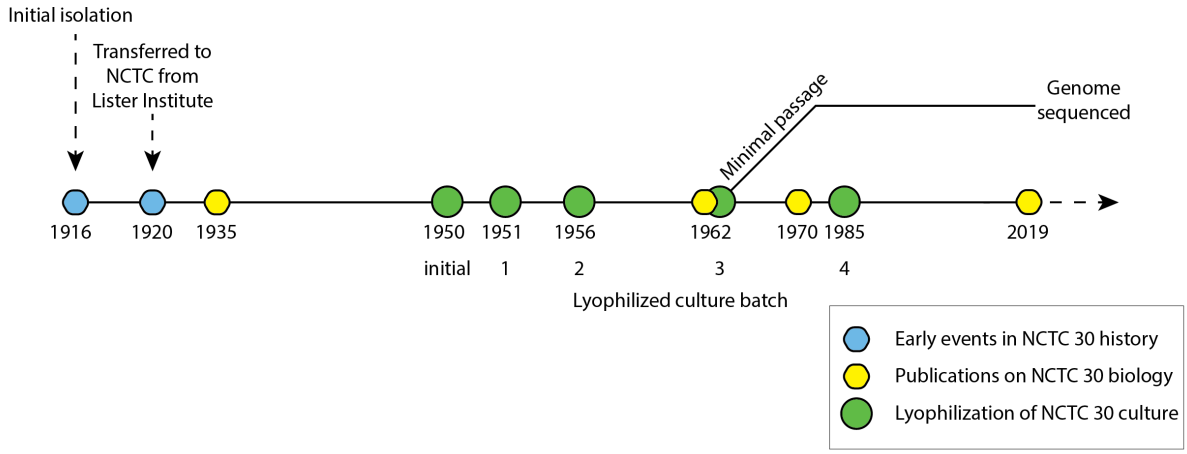


Figure 4.12 – A timeline of events in the history of NCTC 30. Figure based on data contained in Figure 4.11 and reproduced from [425]. Publications on NCTC 30 that are discussed throughout this chapter have been listed (not exhaustive).

gDNA extracted from one clone of minimally-passaged NCTC 30 (MJD382) was sequenced using both short- and long-read technologies (Illumina HiSeq X10 and PacBio RSII). The long-read assembly strategy used for these data (see Methods) produced two circular contigs, and the long reads covered the finished assembly to a depth of 148.01 X. The NCTC 30 genome assembly consisted of two circularised contigs, one corresponding to the larger chromosome of 2,922,904 bases, and one to the smaller chromosome of 1,029,451 bases (Figure 4.13). To guard against genetic rearrangements or contamination introduced through long term storage, a technical replicate extraction of gDNA from MJD382, as well as a biological replicate (gDNA from MJD439, a second colony of NCTC 30 obtained from the first passage of NCTC 30 after rehydration) were also sequenced using both long- and short-read sequencing technologies.

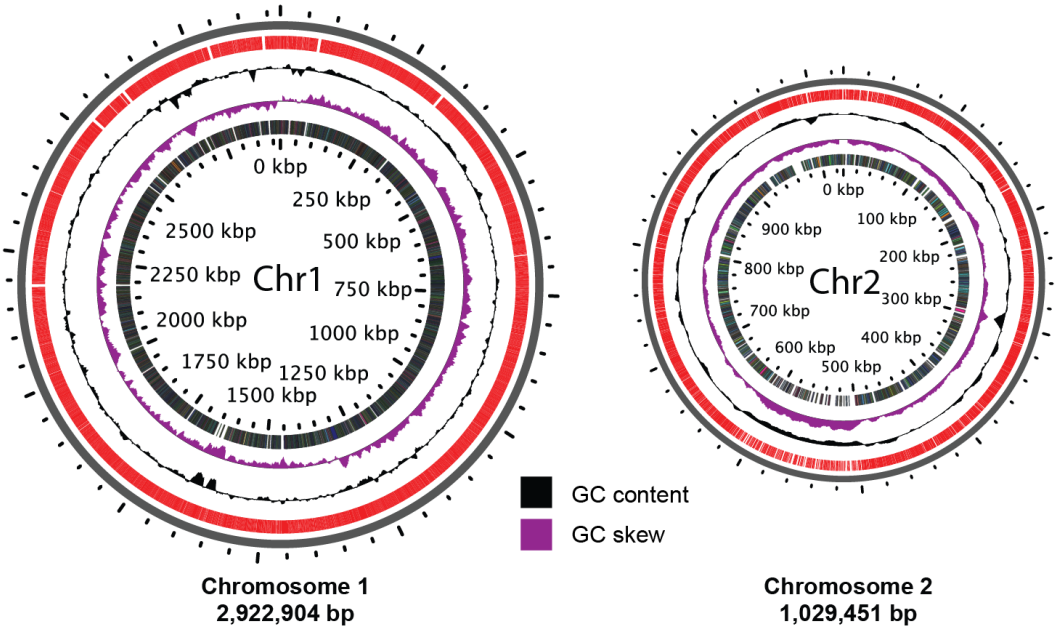


Figure 4.13 – Illustration of the NCTC 30 chromosome sequences. Modified from [424].

A comparison between the NCTC 30 assembly and that of the N16961 reference strain revealed a large inversion in NCTC 30 chromosome 1 of ~ 1,040,746 bases, between genes *VC_1056* and *VC_2013* (Figure 4.14). The short-reads from the same NCTC 30 gDNA preparation used for the assembled long-read sequencing were mapped back to both the PacBio assembly and to N16961, and to guard against this inversion being an artefact of genome assembly using long-reads only, paired reads were identified that mapped to either side of the inversion junction, and individual reads were identified which spanned the junction itself (Figure 4.15). This was confirmed further using sequencing data from a second gDNA isolation from MJD382, as well as from MJD439, a culture grown from a colony of NCTC 30 that had been separated from MJD382 at passage 1 (Methods, section 2.2.4; data not shown).

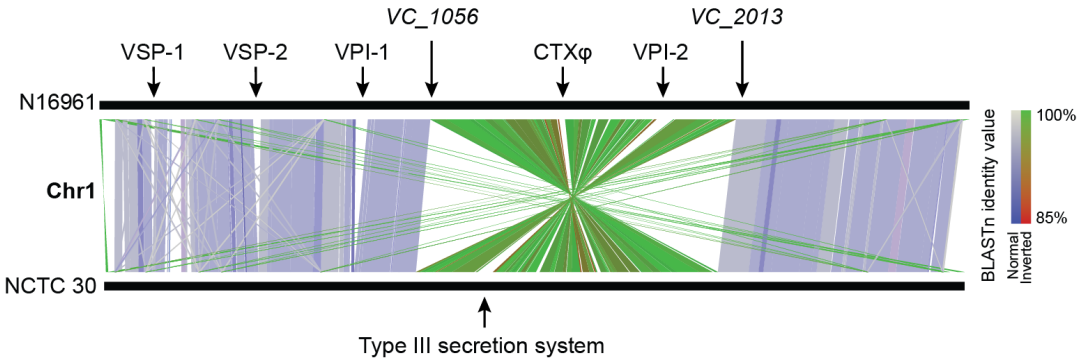


Figure 4.14 – A comparison between chromosome 1 of NCTC 30 and N16961. The chromosomal location of genomic islands and virulence determinants, discussed in subsequent sections, are indicated. The position of the inversion is indicted by the green/red inverted regions of synteny, in an 'hourglass' shape. Modified from [424].

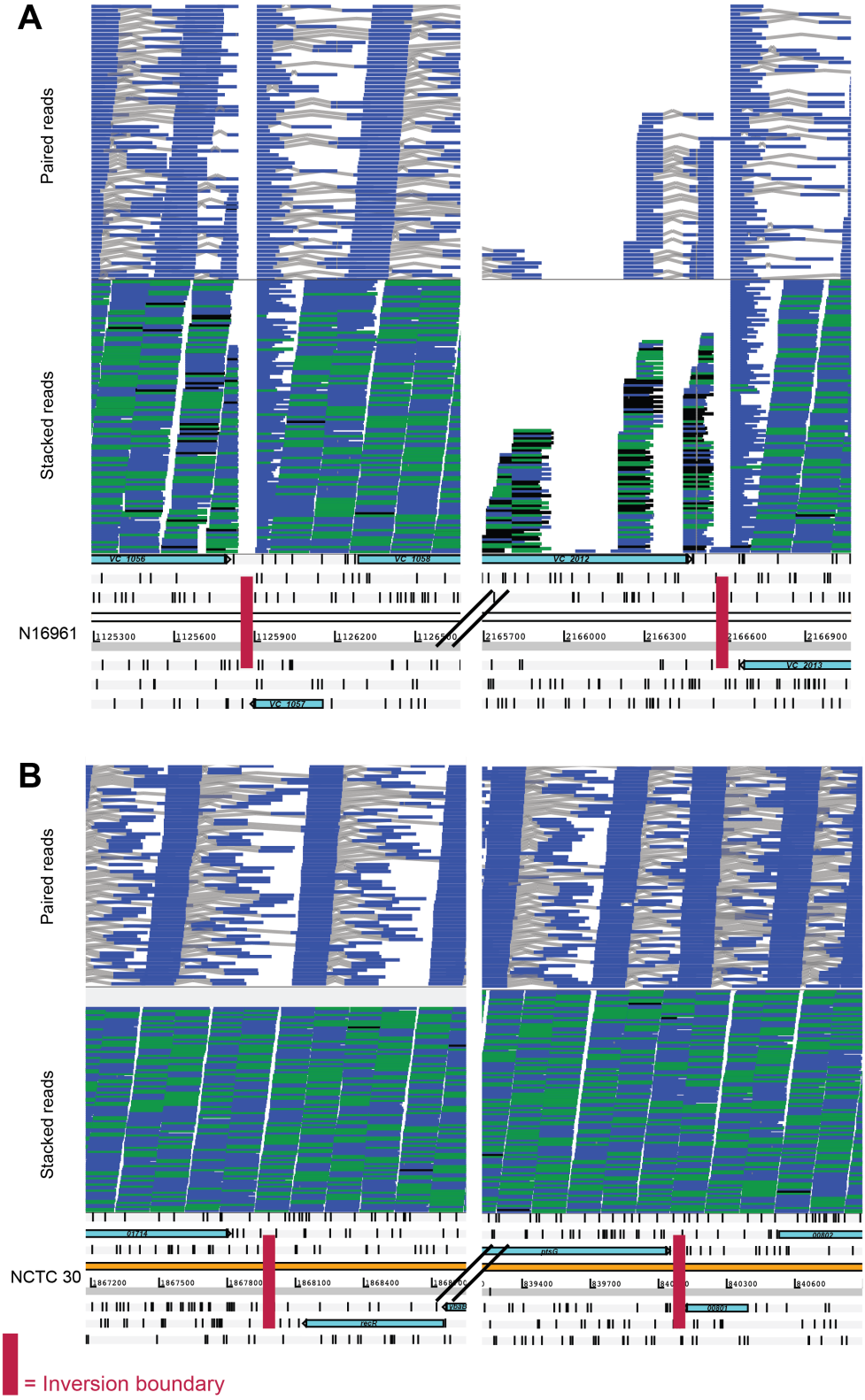


Figure 4.15 – Mapping reads across the inversion junctions. NCTC 30 short-reads were mapped to the N16961 (A) and NCTC 30 (B) genome sequences. The regions delineating the genomic inversion shown in Figure 4.14 are presented. In N16961, no reads map to the inversion region adjacent to *VC_1057*, and none of the reads that map to one of adjacent sequences have a paired read on the other side of the inversion site (A). The inversion site at *VC_2013* is similar. However, when NCTC 30 reads are mapped to the NCTC 30 long-read assembly, reads

map which span the inversion junction, and reads map to each side of the inversion which are paired to reads on the other side of the inversion site (B). The gene *VC_2012* is poorly conserved between NCTC 30 and N16961; hence, few NCTC 30 reads map to this gene. Hatch marks denote a truncation of the genome sequence view.

4.3.8 – NCTC 30 motility and flagellation defects

When NCTC 30 was being revived and cultured, it grew noticeably more slowly than other *V. cholerae* on both solid and liquid media, where it showed erratic growth kinetics (Figure 4.16). The inversion on chromosome 1 was confirmed not to encompass the *crtS* locus, and therefore should not interfere with the rate and timing of chromosome 2 replication, or with the growth rate of NCTC 30 [58] (section 1.2.1).

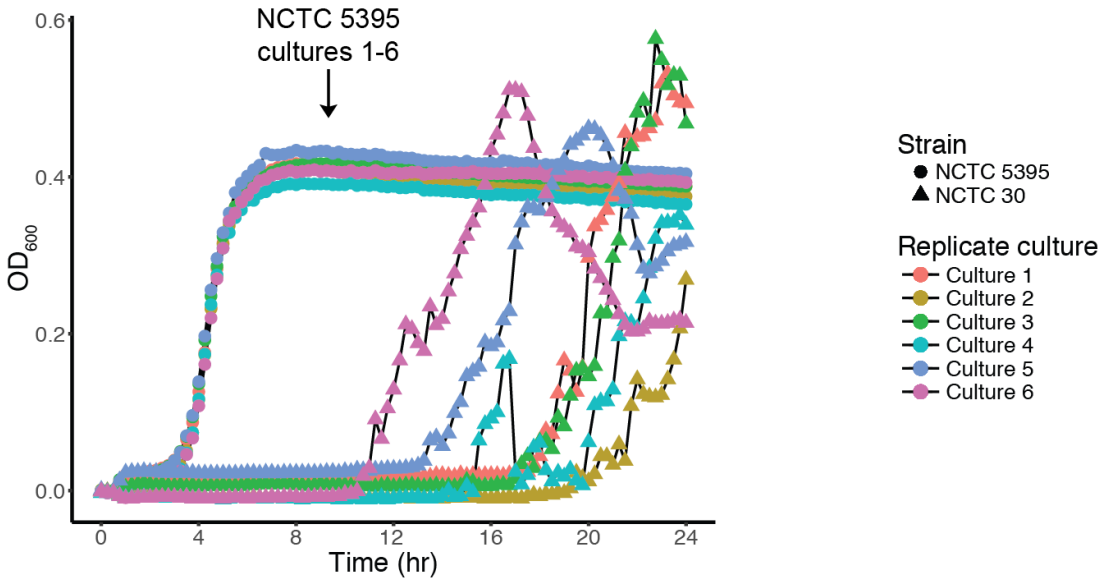


Figure 4.16 – Growth kinetics of NCTC 30 and NCTC 5395. In this assay, NCTC 30 demonstrated a growth defect at 37 ˚C relative to NCTC 5395. Under these conditions, *V. cholerae* does not grow to an OD_{600} exceeding 1.0 – accordingly, a non-logarithmic Y-axis scale has been deliberately used. Representative data from single biological experiments are reported.

This observation prompted the microscopic examination of NCTC 30, using transmission electron microscopy (TEM). Representative TEM images for NCTC 30 and a control strain of classical biotype *V. cholerae*, NCTC 10732, were captured and are presented in Figure 4.17.

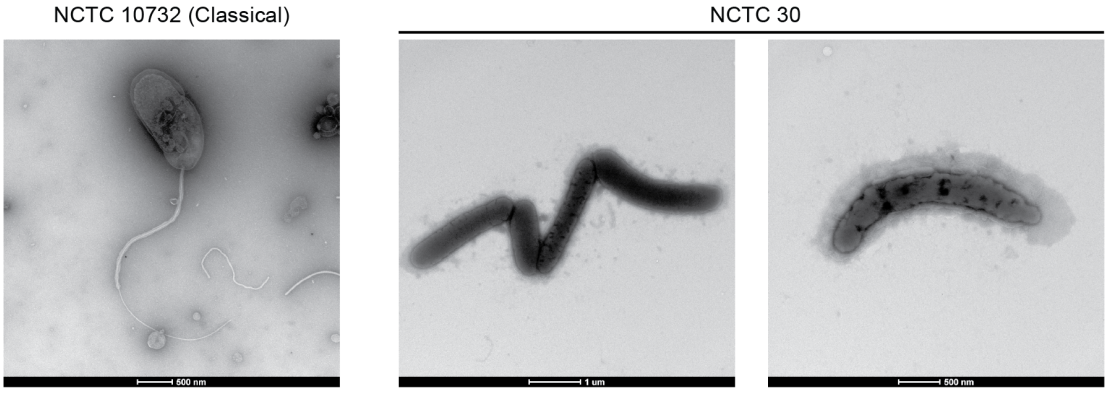


Figure 4.17 – *V. cholerae* transmission electron micrographs. NCTC 30 is compared to a control strain of flagellated *V. cholerae*, NCTC 10732. Modified from [424].

It was noted immediately that NCTC 30 cells did not appear to express the monotrichous flagellum characteristic of *V. cholerae* [1] (Figure 4.17). This was particularly intriguing because previous studies of NCTC 30 had recorded this bacterium as being flagellated [39]. Based on these microscopy data, it was hypothesised that NCTC 30 would be non-motile. Accordingly, NCTC 30 and two control strains of *V. cholerae* were cultured on motility media. No motility was observed in NCTC 30 under these conditions (Figure 4.18).

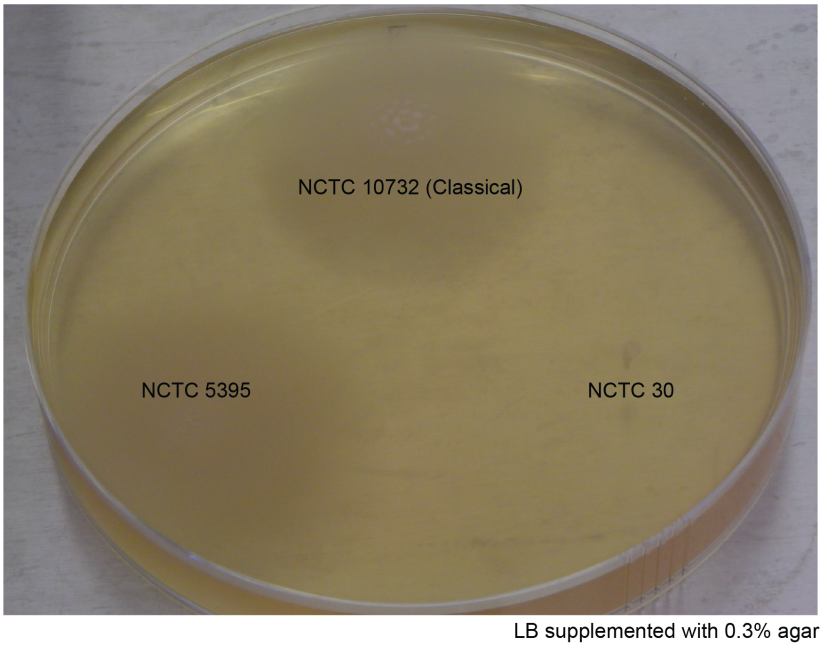


Figure 4.18 – NCTC 30 is non-motile when grown in soft agar. A lateral zone of growth corresponds to motile bacteria 'swimming' within and across the agar; such lateral growth was not observed in NCTC 30. Control strains NCTC 5395 and NCTC 10732 were used for comparisons elsewhere in this chapter. Soft agar plates were incubated for 18 hours before imaging. Reproduced from [424].

Since the flagellar regulatory and biosynthesis hierarchy has been well-studied in *V. cholerae* [426–428] (Figure 4.19), it was decided to attempt to explain this phenotype using the NCTC 30 genome sequence. Each of the genes listed in Figure 4.19 were inspected manually in the NCTC 30 genome assembly, and a putative frameshift was identified in the *flrC* gene (Figure 4.20).

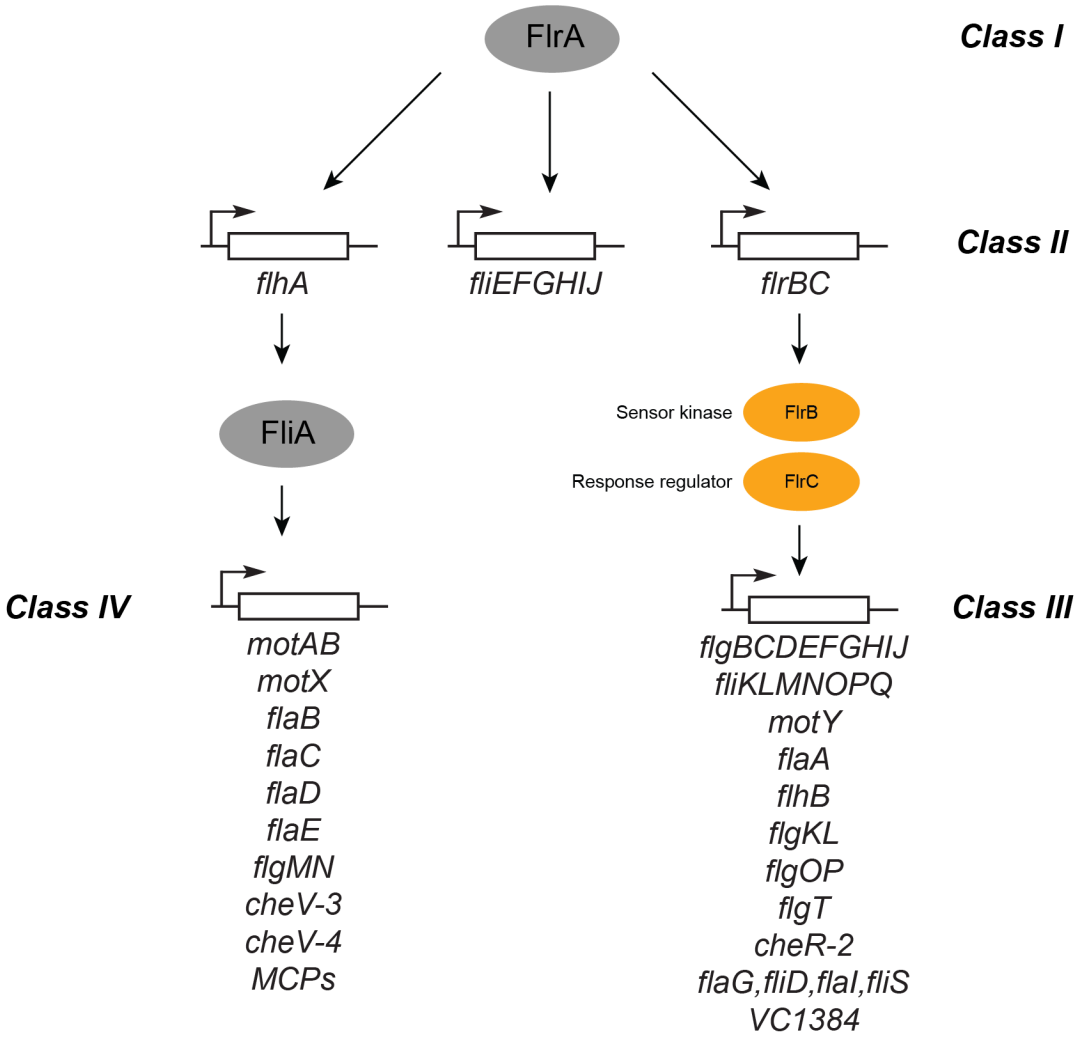


Figure 4.19 – Illustration of flagellar biosynthesis and regulatory hierarchy for *V. cholerae*. The FlrA protein activates transcription of Class II genes, including the *flrBC* operon. FlrB and FlrC then activate multiple Class III flagellar genes. Figure derived from [428].

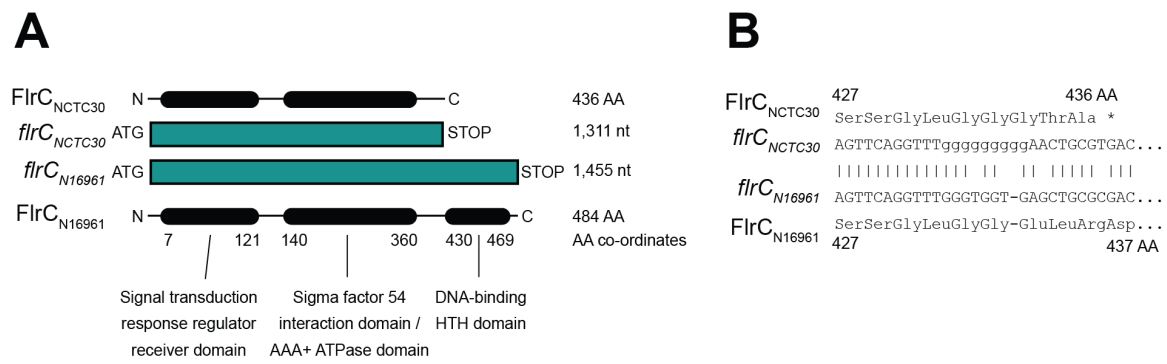


Figure 4.20 – Schematic of the predicted truncation of FlrC caused by the frameshift in *flrC*. (A): The frameshift in *flrC* is predicted to truncate FlrC, removing the DNA binding domain from this protein, thereby eliminating its ability to regulate its target genes. (B): This frameshift is predicted to disrupt the reading frame such that a premature STOP codon is introduced at amino acid position 436. Modified from [424].

The FlrC protein is a response regulator that governs the expression of Class III flagellar genes in *V. cholerae* [426] (Figure 4.19). Class III genes encode components of flagellin, as well as proteins that form part of motor and chemotaxis machinery in *V. cholerae* [427] (Figure 4.19). The mutation detected in *flrC* was predicted to introduce a frameshift, truncating FlrC by removing the last 48 amino acids from the protein (Figure 4.20). This region of FlrC is predicted to encode the DNA-binding domain, and its removal should prevent the transcription of FlrC-regulated Class III genes required for the production of *V. cholerae* flagella.

Although the hypothesis that the *flrC* frameshift was responsible for the absence of flagella was compelling, two additional sequencing technologies were used to ensure that this frameshift was not an artefact of long-read assembly. High-accuracy short-read Illumina data were mapped to the NCTC 30 assembly, and the region was also amplified and sequenced using commercial Sanger sequencing. All three sequencing technologies confirmed the existence of this frameshift in this gene (Figure 4.21).

Additionally, since Davis and Park's previous report was submitted for publication in April 1962 [39], prior to the preparation of batch 3 of NCTC 30 (Figure 4.11), it was necessary to consider the possibility that the *flrC* mutation may have arisen during the preparation of batch 3, during long-term storage [429], or during passage in our laboratory. Since the mutation was confirmed to be present in sequences of MJD382 and MJD439, two independently-sequenced clones separated from each other immediately upon rehydration of our vial of batch 3 NCTC

30, this suggested either that this mutation predated the introduction of the strain into our laboratory, or had arisen immediately upon rehydration of our lyophilised stock.

In order to test whether the *flrC* mutation was unique to our laboratory stocks, gDNA was prepared from batch 4 of NCTC 30 in laboratories at PHE, separate to our laboratory. Batch 4 was lyophilised in 1985 from a culture of batch 3 bacteria (Figure 4.11, 4.12). *flrC* was amplified from batch 4 gDNA by PCR, and Sanger sequencing was used to confirm that the *flrC* frameshift mutation was also present in batch 4 of NCTC 30 (Figure 4.21). This indicates strongly that the mutation arose either during or prior to the preparation of batch 3 of this lyophilised culture, and that this mutation ought to be present in NCTC 30 cultures which are purchased from NCTC in the future.

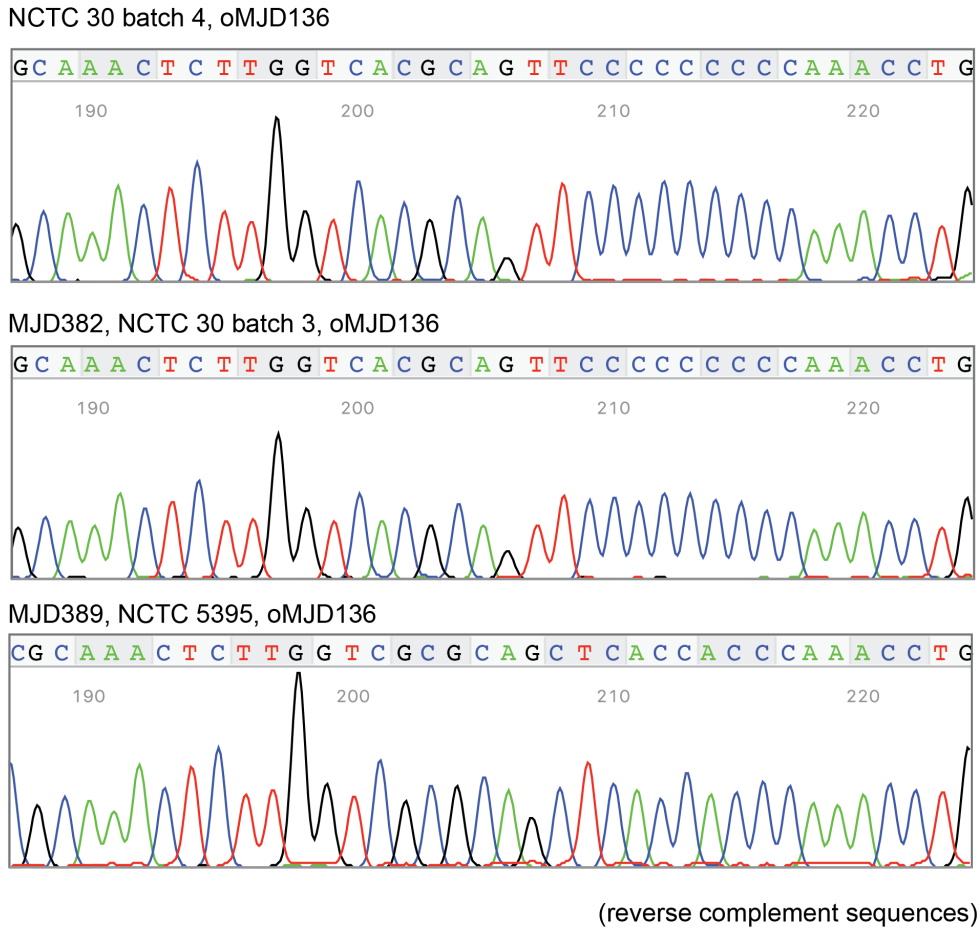


Figure 4.21 – Confirmation of *flrC* frameshift using Sanger sequencing. The mutation that causes the predicted frameshift in *flrC* in NCTC 30 batch 3 (nine G-C base pairs) are also present in the batch 4 DNA preparation. Sequencing traces were visualised using 4Peaks [14].

[14] https://nucleobytes.com/4peaks/index.html

Unfortunately, it proved impossible to culture NCTC 30 efficiently enough to attempt to transform it with plasmid vectors, either to repair the *flrC* mutation by homologous recombination, or to supply an episomal copy of *flrC in trans* to complement the defect. However, it was noted that the microscopic images of NCTC 30 were consistent with that of an *flrB* targeted mutant [426]. FlrB and FlrC form a two-component system in which FlrB phosphorylates FlrC, which then acts as a response regulator and *trans*-activates Class III genes in a σ^{54}-dependent manner [426, 430] (Figure 4.19). Given that *flrC* is the only flagellum-biosynthesis gene in NCTC 30 that is disrupted, it is reasonable to infer that this *flrC* mutation is responsible for this phenotype.

4.3.9 – Antimicrobial resistance in NCTC 30

A previous taxonomic study of *Vibrio* bacteria suggested that NCTC 30 was insensitive to a concentration of penicillin to which a control *V. cholerae* strain, NCTC 5395, was partially sensitive [39]. The NCTC 30 genome assembly was scanned for antimicrobial resistance genes *in silico* and one putative resistance gene, encoding a β-lactamase, was identified. The gene was located within the integron on the smaller chromosome, which had been fully-assembled by virtue of the long-reads. The translated sequence of this gene was used to query the NCBI nr database. The most similar sequences (99% amino acid similarity) were those of the *V. cholerae* β-lactamases CARB-7 and CARB-9. CARB-7 was first described in an environmental *V. cholerae* isolated in Argentina that resisted ampicillin to an MIC of 256 μg/ml, and was also encoded by a gene within the integron of chromosome 2 [152]. CARB-9 is also an integron-encoded β-lactamase first identified in environmental non-O1/O139 *V. cholerae* from Argentina which resisted ampicillin to an MIC of 64 μg/ml [151].

The presence of a gene predicted to encode an antimicrobial resistance determinant neither guarantees that the gene is responsible for a resistance phenotype, nor that it is actually expressed. In order to validate whether this gene was at all functional or responsible for the penicillin insensitivity reported previously [39], within the constraints of working at CL3 and with a difficult-to-culture NCTC 30 (section 4.3.8), it was decided to clone the gene into pACYC184, a low-copy vector encoding resistance to chloramphenicol and tetracycline [359]. The $bla_{CARB\text{-}like}$ gene was amplified from NCTC 30 template DNA. The amplicon was digested

and ligated into pACYC184, disrupting the native *tet* gene on the plasmid (Figure 4.22; Methods, section 2.2.12).

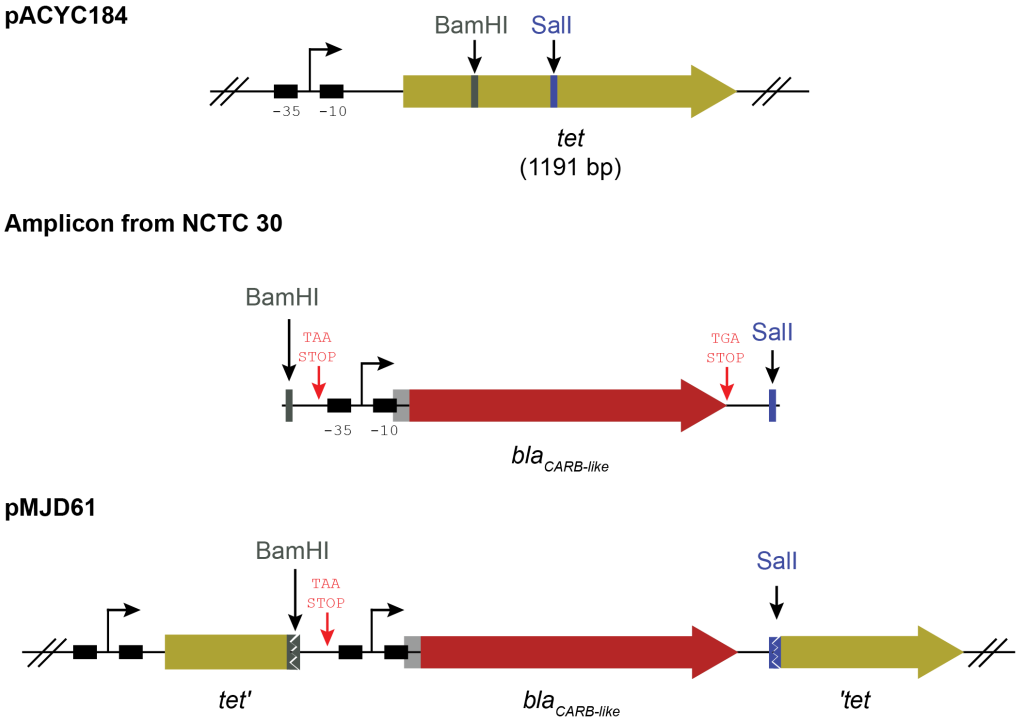


Figure 4.22 – Strategy to clone *bla*$_{CARB\text{-}like}$ from NCTC 30 gDNA. *bla*$_{CARB\text{-}like}$ was amplified from the NCTC 30 genome using primers oMJD96 and oMJD97, incorporating BamHI and SalI restriction sites. This was digested and ligated into the *tet* gene on pACYC184, introducing a premature in-frame STOP codon into *tet*. BPROM [15] predicted *E. coli* σ^{70} -35 and -10 elements within the insert (indicated). Although this software predicts promoters from *E. coli*, not *V. cholerae*, it should be noted that this might provide a native promoter from which *bla*$_{CARB\text{-}like}$ expression may be driven in *E. coli*. Figures are not to scale. Reproduced from [424].

The plasmid harbouring *bla*$_{CARB\text{-}like}$, pMJD61, was transformed into a cloning strain of *E. coli* and MICEvaluator strips were used to compare the relative sensitivities of *V. cholerae* and transformed *E. coli* (see section 2.2.13 for assay details). This assay was not used to determine quantitative MICs. Cells harbouring pUC19 which confers resistance to β-lactams, and pACYC184, an empty vector, were used as positive and negative controls respectively. NCTC 30 was found to be sensitive to ampicillin to a lesser extent than NCTC 5395, the same strain used in Davis and Park's original work [39] (Figure 4.23). The faint growth of NCTC 30 close to the test strip above the 16 µg/ml position resembles satellite colonies that emerge due to β-lactam degradation by enzyme secreted by adjacent bacterial culture (Figure 4.23).

[15] http://softberry.com

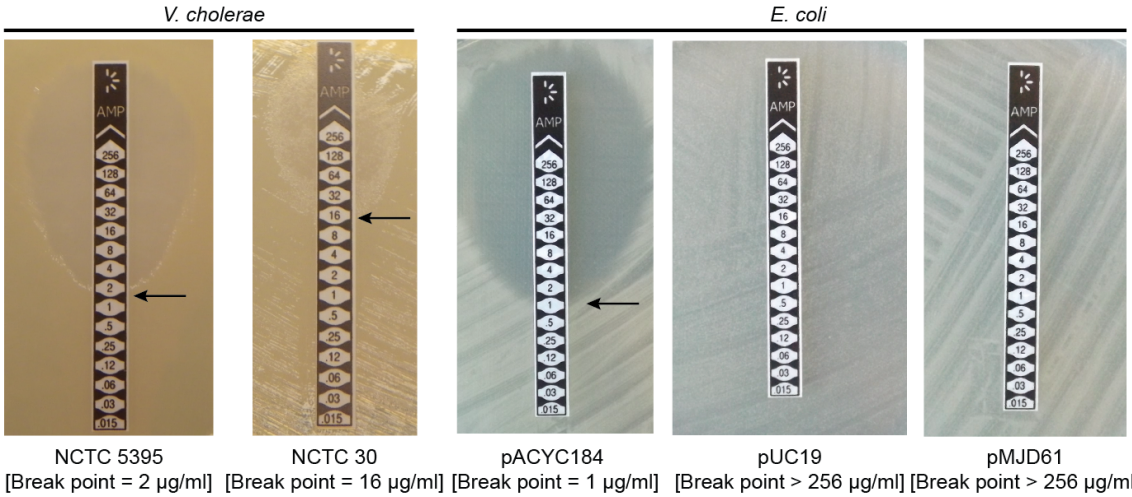


Figure 4.23 – Ampicillin sensitivity phenotypes of *V. cholerae* and plasmid-harbouring *E. coli*. Arrows indicate the break points as interpreted using the MICEvaluator manufacturer's instructions. pMJD61, the plasmid containing $bla_{CARB\text{-}like}$ (Figure 4.22), confers an equivalent ampicillin resistance to that conferred by the pUC19 ampicillin-resistance plasmid in *E. coli* 5-alpha. Modified from [424].

4.3.10 – Virulence determinants in NCTC 30

Earlier comparisons (Figure 4.14) had demonstrated the absence of the canonical pathogenicity islands VPI-1, VPI-2, VSP-1 and VSP-2, as well as the CTXφ prophage from the NCTC 30 genome. Therefore, the sequence was interrogated to attempt to identify genetic determinants that may explain the clinical symptoms giving rise to its isolation (section 4.3.7). The presence of accessory virulence genes in the assembly, including *zot*, *ace*, *hlyA*, *rtxA*, *rtxC*, *hapA*, MSHA and heat-stable enterotoxin, was determined (Table 4.4).

Accessory virulence gene	Locus ID in N16961 reference genome	Present in NCTC 30 (Percentage identity of translated protein)
Zona occludens toxin (Zot)	*VC_1458*	No (CTXφ)
Accessory cholera enterotoxin (Ace)	*VC_1459*	No (CTXφ)
Haemolysin (*hlyA*)	*VC_A0219*	Yes (98%)
Mannose-sensitive haemagglutinin (MSHA)	*VC_0398..VC_0414*	Yes, see Figure 4.24D
MARTX toxin (*rtxA*)	*VC_1451*	Yes (93%)
MARTX toxin accessory gene (*rtxC*)	*VC_1450*	Yes (100%)
HA/protease (*hapA*)	*VC_A0865*	Yes (98%)
Integrative conjugative element SXT/R391	Absent (if present, integrates into *VC_0659*)	No, *VC_0659* homologue is intact
Heat-stable enterotoxin NAG-ST (Genbank accession M85198.1)	Absent	No

Table 4.4 – Accessory virulence genes present in NCTC 30. Identity percentages were calculated by alignment of protein sequences from NCTC 30 and N16961 using BLASTp. The translated NAG-ST nucleotide sequence (accession M85198.1) was used as a tBLASTx query to scan NCTC 30 for the gene encoding this enterotoxin. Modified from [424].

Although it appeared that NCTC 30 was devoid of canonical *V. cholerae* virulence genes (Table 4.4; Figures 4.14; 4.24), it remained the case that this bacterium was of clinical origin – it had been isolated from a convalescent soldier (Figure 4.11), and it had been reported that the source of NCTC 30 was 'choleraic diarrhoea' [41]. Accordingly, the NCTC 30 genome was scanned for other putative virulence determinants seen in other *Vibrio*.

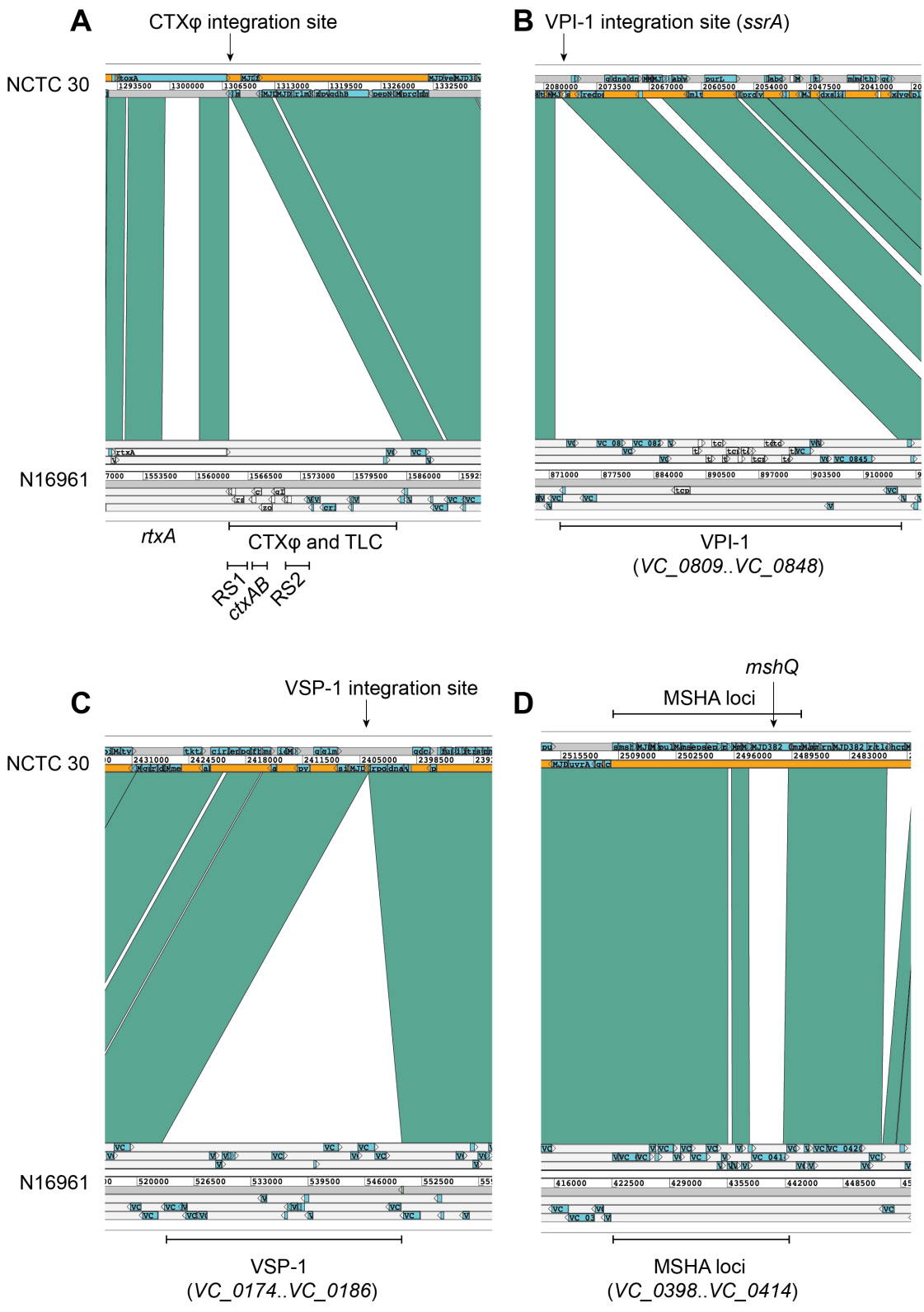


Figure 4.24 – Presence and absence of genomic islands and virulence genes in the NCTC 30 genome assembly. Comparisons are relative to the N16961 reference sequence and compare nucleotide identity. CTXφ (A), VPI-1 (B) and VSP-1 (C) are absent from NCTC 30. The genes that encode the MSHA accessory virulence determinant are present in NCTC 30 (D), although the NCTC 30 *mshQ* gene is dissimilar to that of N16961 (Table 4.4). Reproduced from [424].

A type III secretion system (T3SS) was detected on the larger chromosome of NCTC 30 (Figures 4.14; 4.25). This island is integrated between *VC_1757* and *VC_1810*, the same integration site as used by VPI-2 in N16961 (Figure 4.14, 4.25). This T3SS is more similar to the T3SS found in the genome of *V. parahaemolyticus* strain 10329 [52] than the T3SS found in *V. cholerae* AM_19226, the strain used to characterise T3SS activity in *V. cholerae* [45,46] (Figure 4.25). It is notable that a handwritten note on the NCTC's internal quality check card for NCTC 30 refers to "intermediate *V. cholerae* / *V. parahaemolyticus*" (Figure 4.11). No further information is available to explain why this note was made.

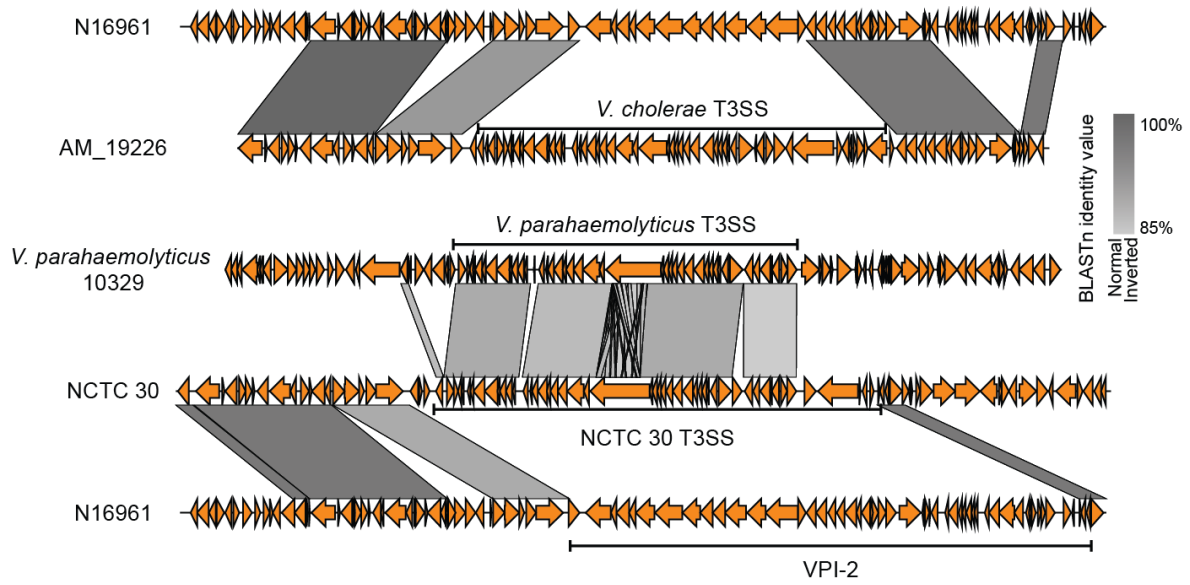


Figure 4.25 – Comparison of T3SS from NCTC 30 and other Vibrios. The T3SS harboured by NCTC 30 is most similar to one found in *V. parahaemolyticus* strain 10329, and is dissimilar to that encoded by *V. cholerae* AM_19226 [266]. The integration locus for T3SS in both NCTC 30 and AM_19226 is the same, and is the same as the VPI-2 integration locus in N16961. However, the genes flanking the T3SS in *V. parahaemolyticus* are not similar to those of *V. cholerae*. Modified from [424].

4.3.11 – Phylogenetic position of NCTC 30 and distribution of T3SS-2β

The T3SS detected in NCTC 30 and *V. parahaemolyticus* 10329 corresponds to the T3SS-2β described in limited numbers of previously-published *V. cholerae* [397] and also found in a number of the Argentinian non-O1 *V. cholerae* described in Chapter 3 (Figure 3.21). Building on the data from Chapter 3, a pangenome was constructed using 197 other *V. cholerae* genome sequences, and those of three *Vibrio* spp. that are closely related to *V. cholerae*. These context genomes are the same as those used to contextualise Argentinian non-7PET *V. cholerae* in

Chapter 3. A maximum-likelihood phylogeny was then calculated using a core-gene alignment of 2,622 genes from this pangenome. NCTC 30 was found to be more closely related to *Vibrio cholerae* than to other members of the *Vibrio* genus, though it was part of a clade separated from many of the *V. cholerae* in this collection (Figure 4.26). This observation is logical when considered together with a taxonomic study of *V. cholerae* performed in 1970, which questioned whether NCTC 30 is a true member of the *Vibrio cholerae* species [38]. The phylogenetic separation which we observed is likely to reflect the phenotypic and molecular differences that questioned the classification of NCTC 30 [38]. Intriguingly, this cluster of isolates is the same cluster of which 48853_F01, the non-toxigenic *V. cholerae* O139 isolate, was a member (Figure 4.9), and to which several of the Argentinian non-7PET *V. cholerae* belonged (Figure 3.21).

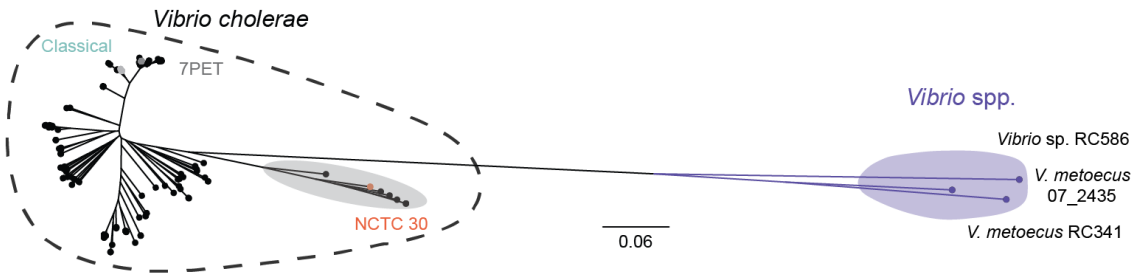


Figure 4.26 – A maximum-likelihood *V. cholerae* phylogeny including NCTC 30. An unrooted phylogeny shows that NCTC 30 clusters together with six isolates that have been previously reported to be *V. cholerae* (grey disc), including 48853_F01, the non-toxigenic *V. cholerae* O139 described in this chapter. The position of 7PET and Classical pandemic lineages [189] are noted. Scale bar denotes the number of mutations *per* variable site. Modified from [424].

Genomes in the pangenome which harboured either T3SS-2β and $bla_{CARB\text{-}like}$ were then identified (BLASTp similarity cut-offs of 95%), to map the distribution of these elements across the phylogeny. Ten genomes harbouring $bla_{CARB\text{-}like}$ homologues were identified both in strains closely related to NCTC 30 as well as in all members of the MX-3 lineage of *V. cholerae* O1, which was isolated in Mexico during 2000 [189] (Figure 4.27). Three *V. cholerae* genomes in the dataset lack CTXφ but contained genes similar to those of the NCTC 30 T3SS-2β [53]. (Figure 4.27).

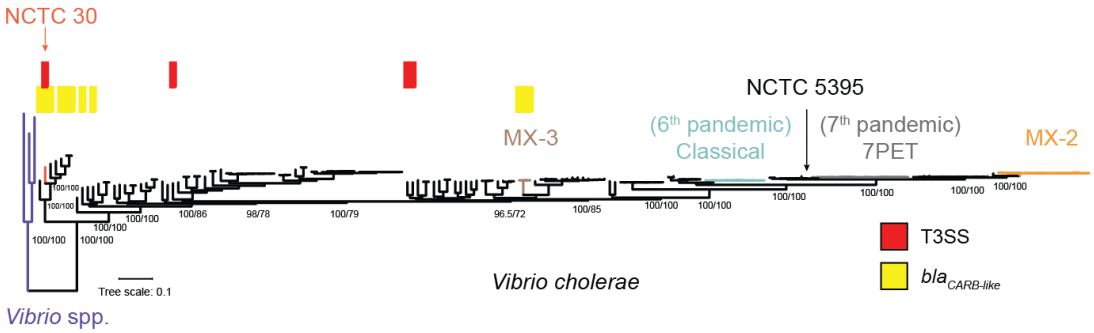


Figure 4.27 – The distribution of T3SS-2β and *bla*$_{CARB\text{-}like}$ within the *V. cholerae* phylogeny. The tree presented in Figure 4.26 is reproduced here, rooted on the three *Vibrio* spp. genomes. Lineages defined previously [189] are indicated, as is the phylogenetic position of the NCTC 5395 control strain used for molecular analyses in this chapter [213]. Modified from [424].

4.4 – Discussion

Piecing together the history of current and previous cholera pandemics requires not only an understanding of pandemic *V. cholerae* lineages, but also a view of the more diverse non-pandemic *V. cholerae* that co-exist contemporaneously with the pandemics. This theme has also been discussed throughout Chapter 3. NCTC 30 was isolated in 1916, at a time when the sixth cholera pandemic was waning [36, 163, 431], but is not a member of the Classical pandemic lineage (Figure 4.26). Very few *V. cholerae* isolates and genome sequences are available from this time period, making NCTC 30 a valuable isolate for future evolutionary studies of the *V. cholerae* species. Similarly, the non-toxigenic *V. cholerae* O139 was isolated during the seventh cholera pandemic, but was not related to the O139 sub-lineage of the 7PET pandemic lineage.

The manual and focused fine-scale analysis of single genomes presented in this chapter is highly complementary to the larger-scale analyses described in Chapter 3. The observations made from the study of these genomes can be extrapolated to larger datasets easily – for example, examining the distribution of T3SS and β-lactamases across larger collections of *V. cholerae* (Figure 4.27). Similarly, although it was by virtue of manual inspection that the coincidence of more than one *ctxB* allele in one closed *V. cholerae* O139 genome assembly was observed, once it was clear that this needed to be looked for, it was feasible to examine larger collections of *V. cholerae* O139 genomes for the coincidence of *ctxB4* and *ctxB5*, including those *V. cholerae* O139 from Domman *et al* [235]. This approach – the extrapolation of knowledge gleaned from the study of single genomes and laboratory strains into the context of larger populations – is a theme that will be re-visited in Chapter 5.

In vitro data strongly support the hypothesis that 7PET *V. cholerae* participated in homologous recombination, mediated by natural competence, to convert from serogroup O1 to O139 [144]. *V. cholerae* O139 have only caused cholera epidemics in South East Asia [244]. It is interesting to speculate that a bacterium similar to the non-toxigenic *V. cholerae* O139 sequenced in this study might have been the source of the O139 operon which was acquired by 7PET in this event. Likewise, the O139 *V. cholerae* that are distantly related to 7PET are likely to have originated from non-O1 progenitors, as suggested previously [230]. It is also evident that non-7PET *V. cholerae* O139 may harbour pathogenicity islands – among the non-7PET serogroup O139 isolates studied by Siriphap *et al* were four non-toxigenic isolates obtained in 2011, all

of which were shown to harbour VPI-2 [242]. The closed toxigenic and non-toxigenic genome sequences reported in this chapter will therefore serve as useful reference sequences for the future genomic analysis of both pandemic and non-pandemic *V. cholerae* O139.

A recurring theme from Chapters 3 and 4 is the presence of T3SS in non-O1 *V. cholerae* of clinical origin (Figure 3.21; Figure 4.27). It may be that the T3SS encoded by NCTC 30 was responsible for clinical symptoms that led to the isolation of these bacteria, not least because it was the most compelling virulence determinant identified in the genome of this bacterium. Although a detailed characterisation of this element was beyond the scope of this PhD project, this is the focus of future study. The possibility cannot also be excluded that the patient was co-infected with another pathogen in addition to NCTC 30, perhaps an O1 *V. cholerae* or another bacterium such as enterotoxigenic *E. coli* [54,55], which might also have caused "choleraic diarrhoea". Although co-infection may explain the isolation of a non-toxigenic *V. cholerae* O139 from a patient suffering from diarrhoea, it is known that even the non-toxigenic vaccine strains of *V. cholerae* O1 can elicit a diarrhoeal response in vaccine recipients [432]. Thus, the contribution of co-infections and non-toxigenic *V. cholerae* to enteric disease should be the focus of further research.

The study of $bla_{CARB\text{-}like}$ in NCTC 30 was valuable because it provided confirmation that this gene can confer ampicillin resistance, and provides genomic explanations for previous phenotypic studies of this isolate [39]. This confirmation was necessary because NCTC 30 predates the introduction of penicillin as an antibiotic, the antimicrobial activity of which was first reported by Fleming in 1929 [433]. Consequently, NCTC 30 is unlikely to have acquired its drug resistance phenotype in response to selective pressures imposed by the therapeutic use of antibiotics. It is reasonable to speculate that NCTC 30 may possess $bla_{CARB\text{-}like}$ in order to resist antibiotics found in its environment – *i.e.*, to defend itself against antibiotic-producing micro-organisms with which it might co-exist in the environment. This might also explain why NCTC 30 appears not to resist the antibiotic completely (Figure 4.23); it may be that $bla_{CARB\text{-}like}$ is expressed at levels sufficient to protect NCTC 30 from diffuse, low-concentration antibiotics present in an environment. It is also notable that β-lactams are not recommended for the treatment of cholera [11, 434], and that although a β-lactamase gene homologous to $bla_{CARB\text{-}like}$ ($bla_{CARB\text{-}2}$) was identified in the MX-3 lineage of *V. cholerae* O1, antimicrobial susceptibility testing (AST) performed on an isolate from this lineage did not lead to this isolate being classified as resistant to β-lactams [189]. This may reflect variety in β-lactam resistance

phenotypes in *V. cholerae*; bla_{CARB-2} might elevate β-lactam resistance in MX-3, but not to a level sufficient to classify a strain as "resistant" to an antimicrobial.

One of the most striking observations in this chapter is that both 48853_F01 and NCTC 30 occupy the same clade of the *V. cholerae* phylogeny, one which is very distinct from the section of the tree which containing pandemic clones and other *V. cholerae* O1 isolates (Figures 4.9, 4.26, 4.27). The phylogenetic separation which was observed is likely to reflect the phenotypic and molecular differences that led to the taxonomic classification of NCTC 30 being questioned [38], though these collective data do strongly indicate that isolates in this clade are indeed *V. cholerae* rather than another species *per se*. The structure and gene content of this clade will be explored in more detail in Chapter 5.

In Chapter 5, I will collate the genomes discussed in this chapter and the non-7PET Argentinian genomes from Chapter 3, together with an additional set of genome sequences from historical and contemporary diverse *V. cholerae*, to produce a collection of ~600 genome sequences. Using these data, I will explore the distribution of key virulence determinants across the diversity of *V. cholerae*, with particular focus on the phylogenetic clade to which 48853_F01 and NCTC 30 belong, on T3SS and accessory virulence determinants, and on the genes and SNVs which determine the El Tor biotype.

Chapter 5

The accessory genome - concordance and conflict between *V. cholerae* genomics and phenotypic dogma

Contribution statement

Nick Thomson supervised the work described in this chapter. NCTC cultures were supplied by Sarah Alexander and Julie Russell (NCTC). Jake Turnbull assisted with the collation of NCTC internal records. Additional isolates were supplied by Florian Marks (IVI) and Claire Jenkins (PHE). The Mexican isolates and metadata described here were supplied by Alejandro Cravioto. Charlotte Tolley extracted gDNA from a limited number of live isolates at WSI under my supervision.

I performed all experiments and analyses, and produced all figures.

COVID-19 statement

The work described in this chapter was affected by the shutdown imposed on the University of Cambridge and the Sanger Institute by the COVID-19 pandemic. This affected planned experiments to validate *in vitro* the genomic observations described in this chapter.

5.1 – Overview

Chapters 3 and 4 have both highlighted the striking differences in diversity between pandemic and non-pandemic *V. cholerae*. Chapter 3 described a focused study of a clonal sub-lineage of pandemic *V. cholerae*, and presented an initial characterisation of 65 non-pandemic *V. cholerae* from Argentina. In Chapter 4, an effort was made to explore the diversity of *V. cholerae* beyond the pandemic lineage by focusing specifically on the forensic analysis of a small number of closed genome sequences, taken from specific *V. cholerae* that were of biological and historical interest.

Much of our understanding of the differences between pandemic and non-pandemic V. cholerae, particularly at the level of gene expression and regulation, come from detailed molecular studies of a handful of reference strains of bacteria. These include 7PET isolates such as N16961, the strain used for the initial *V. cholerae* sequencing study, and C6706 and A1552, both from Latin America [54, 59, 378]. Classical isolates such as O395 [97] and 569B [219, 435] have been well-characterised, as have a handful of serogroup O139 isolates [436] and non-O1 isolates such as V52 and AM_19226 [256, 266]. However, the insights gleaned from these laboratory strains are rarely extrapolated into a wider genomic context, to understand how key regulators of gene expression or phenotypic determinants are distributed and vary across a diverse population of bacteria.

Expanding horizons to consider larger and more diverse collections of genome sequences increases the context into which our understanding can be placed. For example, by studying a single representative isolate of classical biotype *V. cholerae* (O395), it was initially hypothesised that all classical biotype (and thus, Classical lineage) isolates possessed a mutation in *hapR*, preventing HapR-dependent regulation of gene expression in response to cell density [437]. However, considering additional genomes caused this generalisation to be disproven – although O395 does indeed have an inactivating mutation in *hapR*, this is not the case in all Classical isolates, such as CA401[438].

Having access to as many diverse genomes as possible is therefore essential to allow for a maximally-unbiased study of the *V. cholerae* species. For instance, it has been stated previously that in *V. cholerae*, plasmids are rare, and drug resistance is usually encoded by conjugative transposons in the bacterial chromosome [230]. In order to make accurate determinations about

the frequency of plasmid types or antimicrobial resistance determinants across a species, it is essential not to focus exclusively on 7PET – as has been demonstrated in earlier chapters in this thesis, the dynamics of 7PET and non-7PET *V. cholerae* are extremely different.

In this chapter, I present an analysis of a collection of diverse, non-pandemic *V. cholerae*. These were collated from several sources, and include a set of historically-important isolates from the NCTC collections, a set of recent isolates from travellers returning to the United Kingdom [439], and non-O1 isolates of both clinical and environmental origin from Mexico. These were added to the diverse phylogeny of isolates presented in previous chapters, to expand as much as possible the sequenced diversity of *V. cholerae*.

5.2 – Specific aims

In this chapter, I aim to:

1) Determine the population structure of a large collection of genomically-diverse *V. cholerae* isolates,
2) Characterise the distribution of key virulence, antimicrobial resistance genes across the dataset,
3) Characterise the genomes of a number of historically-significant isolates, and use these to understand phenotypic observations made about these isolates in the past, and
4) Explore the molecular basis of biotypes across the *V. cholerae* species.

The isolates described in this chapter were particularly valuable because information on their history and provenance was available for analysis. The dataset consisted of a wide variety of *V. cholerae* collected as part of several different projects. In this chapter, a targeted approach was taken to characterise the distribution across the species of genetic determinants that are specifically associated with important phenotypes.

5.3 – Results

5.3.1 – Expansion of the V. cholerae *species phylogeny*

Two hundred and sixty-eight additional genomes were added to the 383 genomes used to compute the non-7PET phylogenies presented in Chapters 3 and 4. A deliberate effort was made to include non-O1 *V. cholerae* when compiling these genomes, to attempt to maximise the genetic diversity within the dataset. After excluding contaminated and poorly-assembled sequences, a pangenome was calculated using the final collection of 651 sequences (11 isolates were sequenced twice, meaning that 640 independent sequences are represented in this dataset). Using 187,675 variable sites in a core-gene alignment of 2,721 genes, a maximum-likelihood phylogeny for these isolates was calculated. Statistical support for this phylogeny was determined using 5,000 ultrafast bootstrap approximations (Figure 5.1).

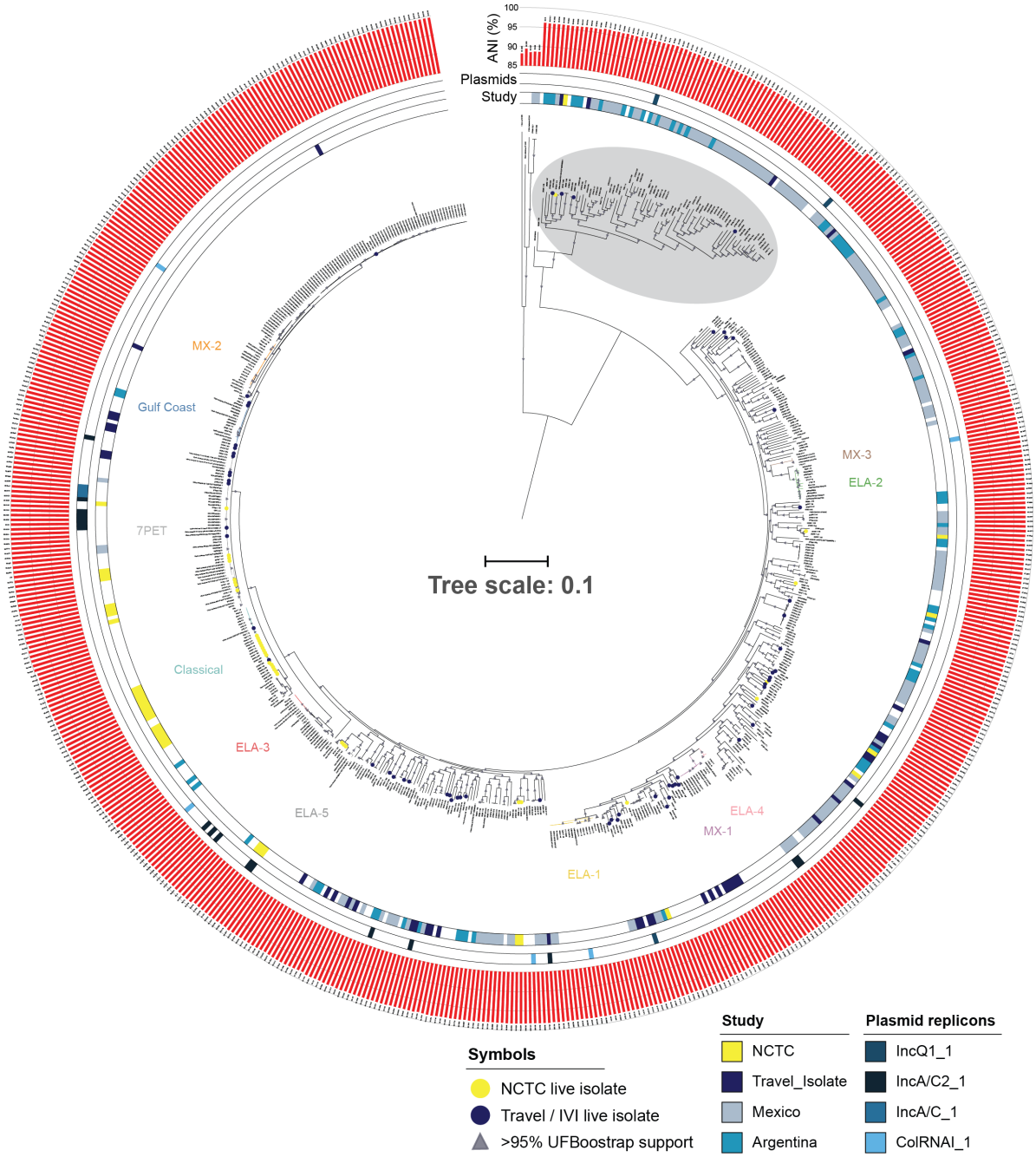


Figure 5.1 – A maximum-likelihood phylogeny of 646 *V. cholerae* and 5 *Vibrio* spp. Triangles on nodes indicate a node that has >95% UFBootstrap support (5,000 replicates). The outer ring denotes the strain collection from which each sequence was obtained. Coloured circles on leaves indicate that a live culture of the sequenced isolate was available at WSI for subsequent experimental analyses. Lineages were coloured and named for consistency amongst thesis chapters. Height of red bars indicates ANI percentage relative to the A1552 reference sequence as described in Chapter 3. Plasmid replicons were identified using ABRicate and the PlasmidFinder database (see below). The tree is rooted on the *Vibrio* spp. outgroup. Scale bar denotes substitutions *per* variable site. The clade of isolates to which NCTC 30, 48853_F01 and multiple Argentinian non-O1/O139 *V. cholerae* belong is highlighted in each case (grey disc).

5.3.2 – Initial characterisation of diverse V. cholerae

It was apparent that two of the Mexican isolates sequenced in this study clustered adjacent to the *Vibrio* spp. outgroup (Figure 5.1). Inspection of the Kraken QC report for these two isolates confirmed their identity to be *Vibrio metoecus*. These two sequences were retained as part of the outgroup for subsequent analysis. A clade of divergent *V. cholerae*, to which NCTC 30 and the non-toxigenic *V. cholerae* O139 (48853_F01) belong, was also substantially expanded as a result of adding these diverse genomes (Figure 5.1).

Comparing the topology of *V. cholerae* species phylogenies from Chapters 3 and 4 to the phylogeny discussed in this chapter provided an intuitive overview of the diversity being discovered as non-O1 isolates were added to the sequence collection (Figure 5.2).

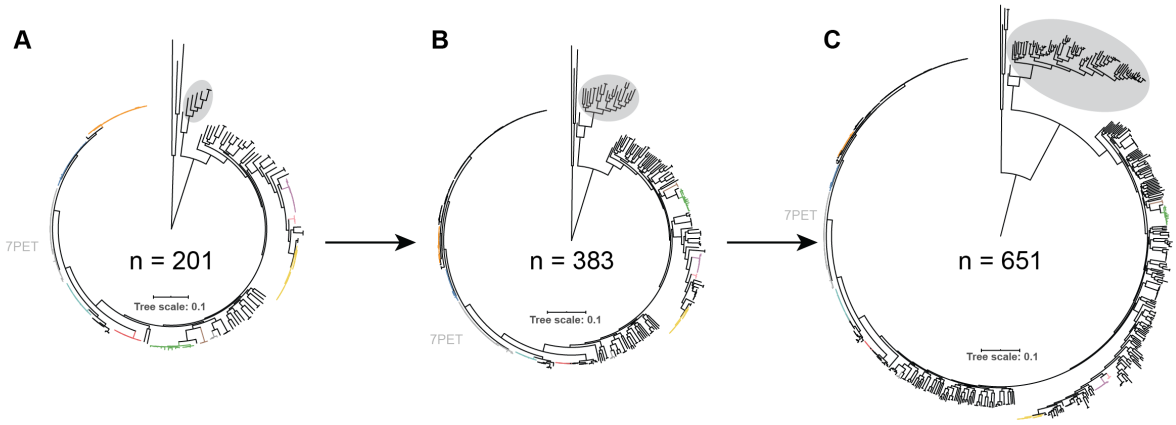


Figure 5.2 – Illustrating the iterative expansion of the *V. cholerae* phylogeny during this thesis research. A: The phylogeny from Figure 4.26, consisting of previously-published isolates and NCTC 30. B: The phylogeny from Figure 3.21, which consists of the dataset used in (A), Argentinian *V. cholerae*, and published Chinese genomes. C: Adding 217 genomes for this chapter to the phylogeny in (B). This phylogeny is presented in detail in Figure 5.1. Lineages are coloured as in Figures 3.21, 4.26 and 5.1. Scale bars denote substitutions *per* variable site. All phylogenies rooted on *Vibrio* spp.

In addition to phylogenetic analysis, average nucleotide identity (ANI) calculations relative to the A1552 reference were performed for each genome assembly, to provide an overview of the relative differences in nucleotide diversity across the phylogenetic tree. These data show that sequences in this divergent clade have a mean ANI of 96.15% relative to A1552 (min 95.85, max 96.42), which is significantly different both to the remaining *V. cholerae* (mean 98.65, min 97.54, max 100) and *Vibrio* spp (mean 88.6, min 88.0985, max 89.3209) (Figure 5.3).

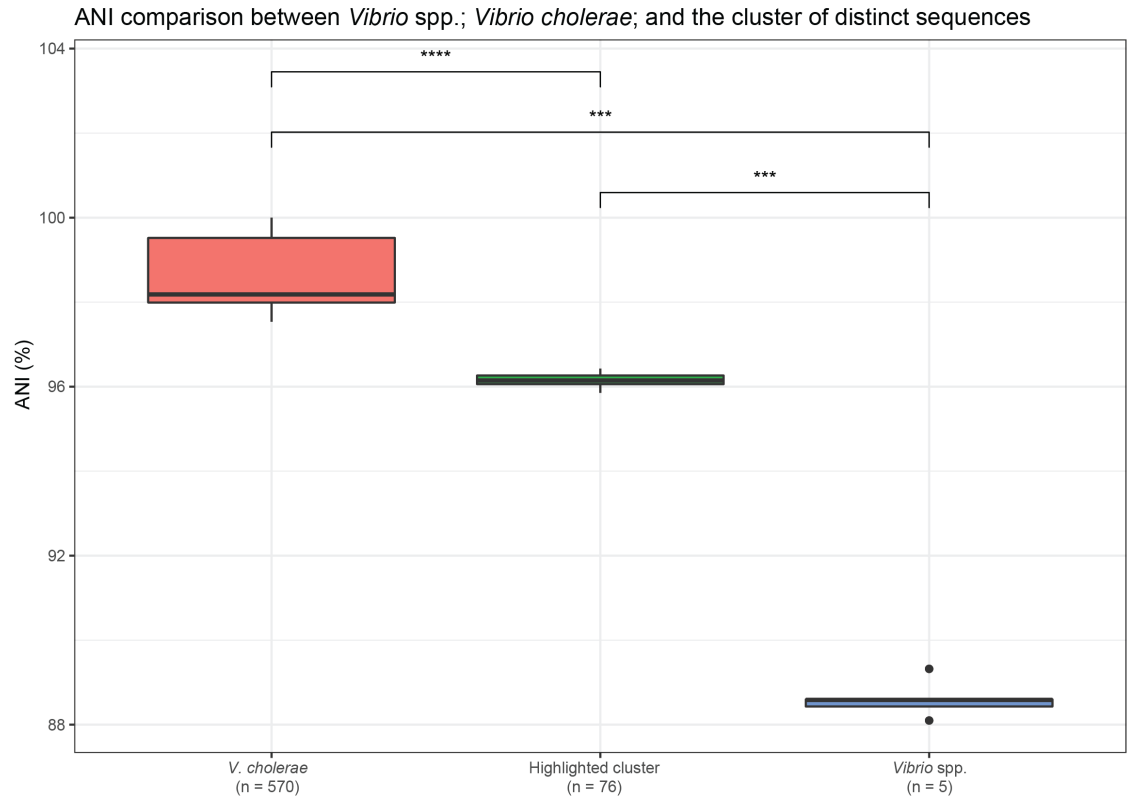


Figure 5.3 – Comparing ANI values for *V. cholerae*, *Vibrio* spp., and the group of diverse sequences highlighted in Figures 5.1 and 5.2. A Wilcoxon rank-sum test was used for statistical testing, to avoid assuming that ANI data were normal in nature. ***: $p < 0.001$, ****: $p < 0.0001$. The highlighted cluster corresponds to isolates contained within the grey disc in Figure 5.1.

The genomic heterogeneity of this divergent clade relative to the remainder of *V. cholerae* in the dataset was also illustrated by the gene presence/absence matrix for this pangenome (Figure 5.4). In this visualisation, clusters of genes absent from this cluster but present in the remainder of sequenced *V. cholerae* are evident.

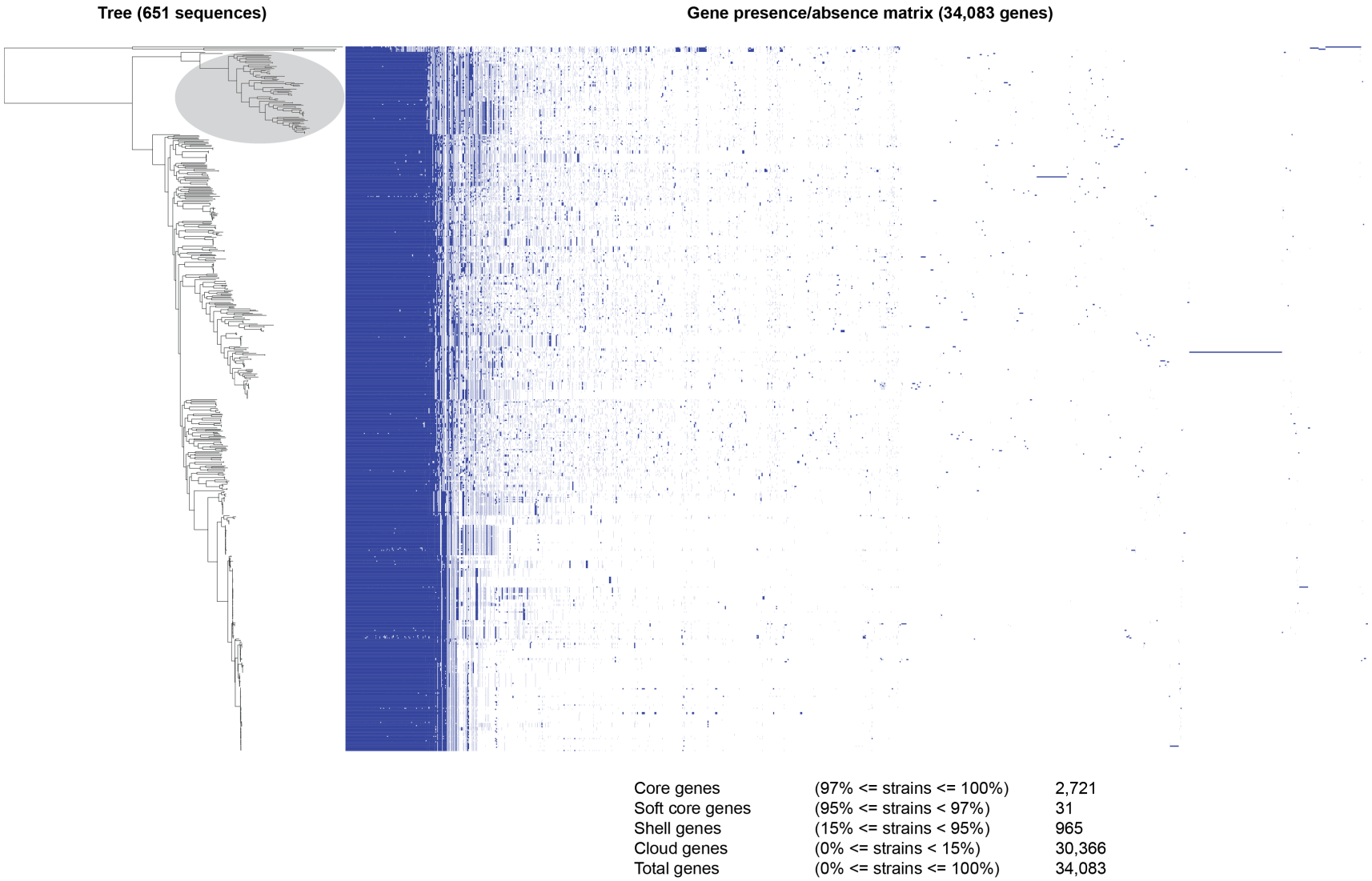
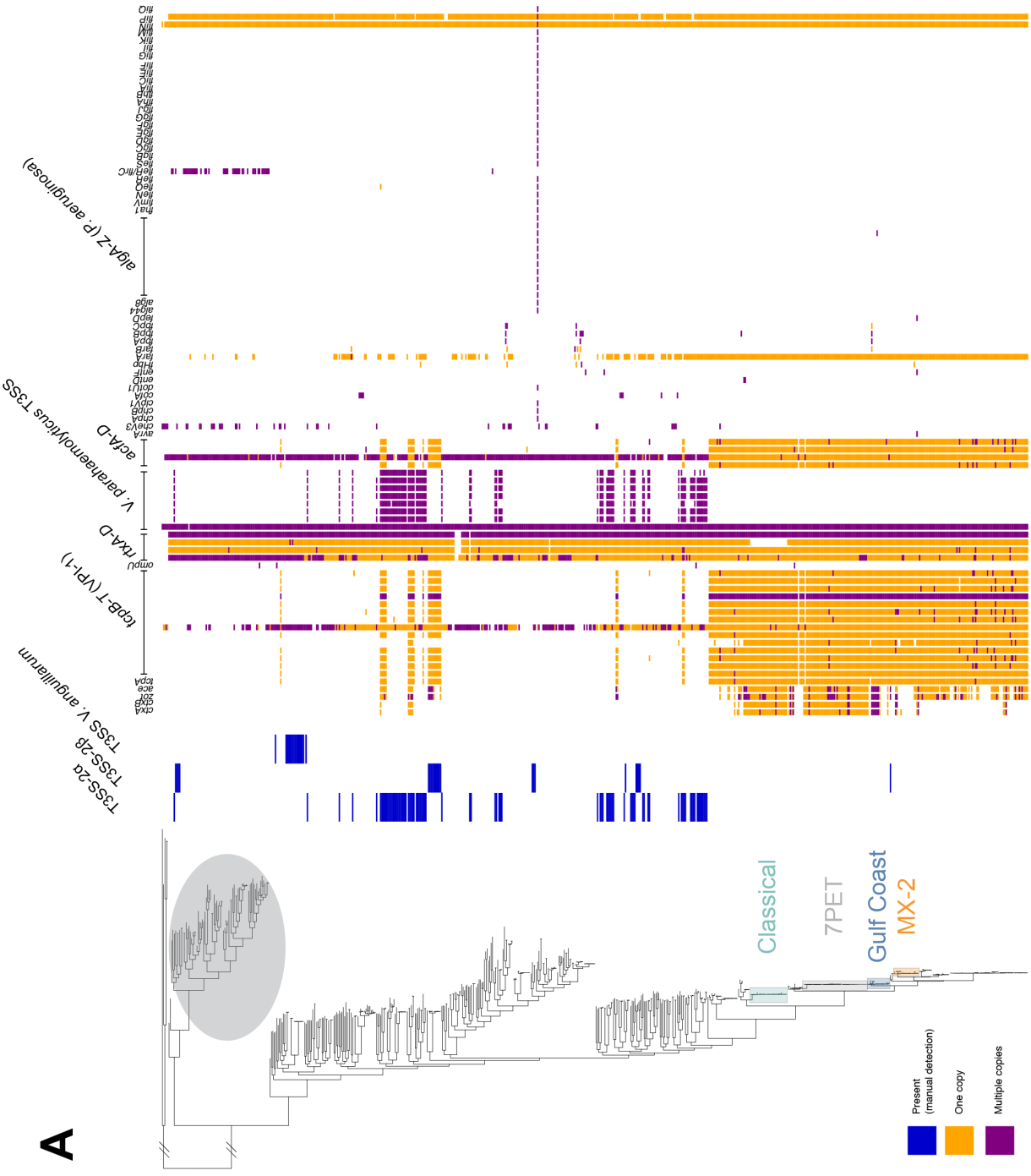


Figure 5.4 – Summary statistics and visualisation of the gene presence/absence matrix for the expanded *V. cholerae* phylogeny. The grey disc indicates the set of diverse *V. cholerae* discussed in section 5.3.2

5.3.3 – Virulence gene distribution across the V. cholerae *phylogeny*

To begin to explore the diversity of these genomes, the distribution of orthologues of known virulence determinants across the dataset was determined. Many of the non-pandemic isolates for which genomes were available were obtained from clinical sources – *i.e.*, they caused an illness in a human patient. Although the clinical metadata for these isolates are sparse, determining the possible mechanisms by which these isolates might have caused illness from their genome sequences was performed. To do this, the presence of genes included in the VFDB database was determined across the pangenome. In addition, the analysis of Argentinian non-7PET genomes in Chapter 3 had identified genes corresponding to three distinct T3SS. These were used to determine whether such systems were present in this expanded dataset using the pangenome gene presence/absence matrix. These results are presented in Figure 5.5.


A
T3SS-2α
T3SS-2β
T3SS V. angullarum
tcpB-T (VPI-1)
tfxA-D
V. parahaemolyticus T3SS
actA-D
aIgA-Z (P. aeruginosa)
Classical
7PET
Gulf Coast
MX-2
Present (manual detection)
One copy
Multiple copies

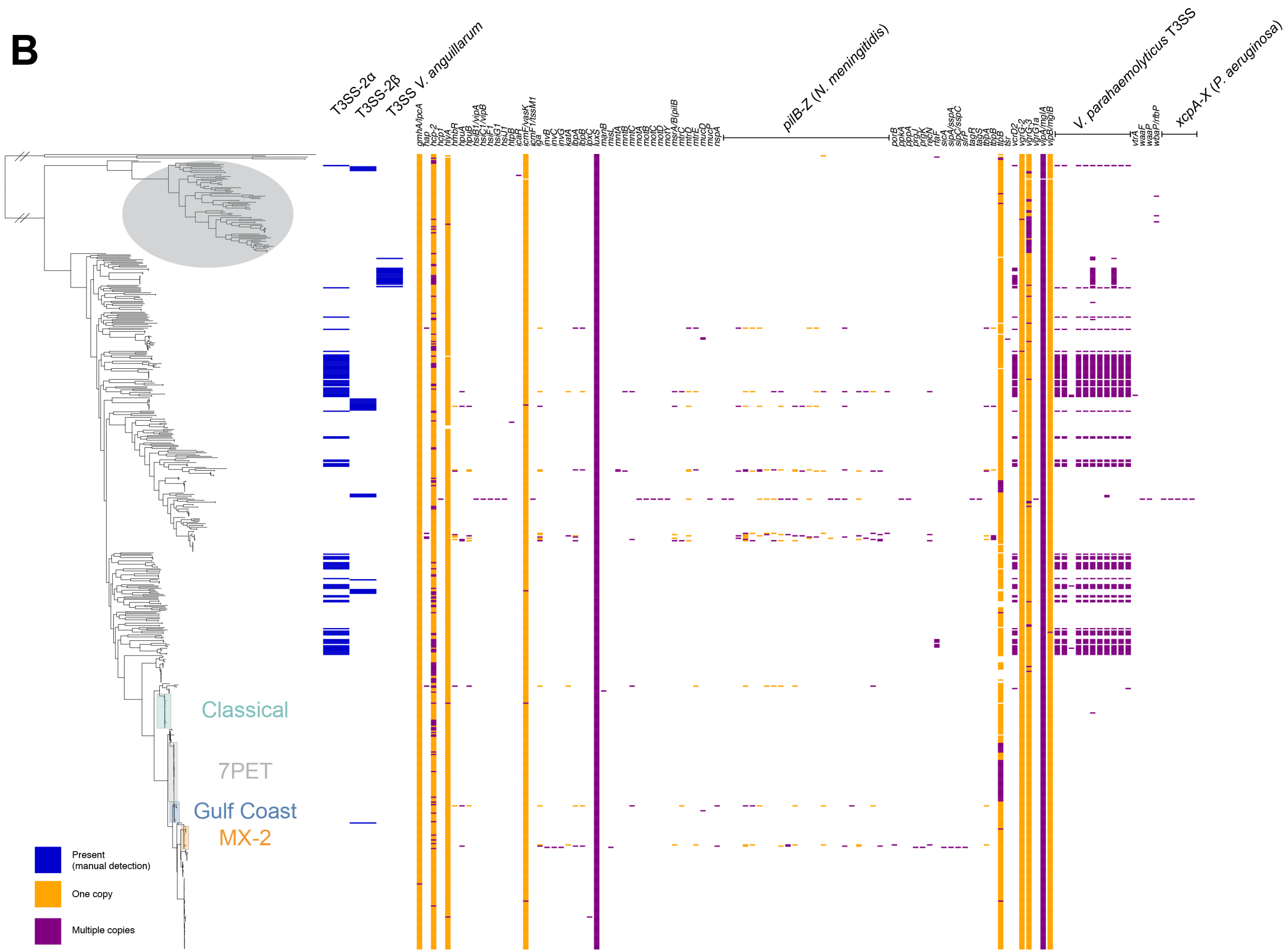


Figure 5.5 – Distribution of key virulence genes across the *V. cholerae* phylogeny. Legend on adjacent page

Figure 5.5 legend: The presence of T3SS elements was determined by identifying genomes containing orthologues of the T3SS genes characterised in Chapter 3. The genome for isolate SRR221551 was the only sequence to contain a large set of *P. aeruginosa* virulence genes. Manual inspection of this assembly and verification of its Kraken report demonstrated that this sequence was contaminated with *Pseudomonas* sequence. Hatch marks denote branches which have been manually shortened for illustrative purposes. Figure has been split into two sub-figures for legibility purposes.

Clusters of T3SS genes appear to be distributed amongst the diverse *V. cholerae* in the dataset, though they are absent from the epidemic lineages and the lineages related to these (Figure 5.5). It also appears that T3SS are rare amongst the clade of highly-diverse isolates, though T3SS-2β was detected in NCTC 30 and related isolates (Figure 5.5). The presence of *V. parahaemolyticus* T3SS genes, as identified by ABRicate using VFDB, appears to indicate the presence of T3SS-2α (ABRicate did not identify complete T3SS in genomes harbouring either the T3SS-2β or the *V. anguillarum* T3SS described in Chapters 3 and 4). However, two of the T3SS-2α genes included in VFDB were detected in isolates harbouring *V. anguillarum* T3SS, suggesting that these genes are conserved amongst or common to the two systems. Genes encoded by VPI-1 were found throughout pandemic lineages, related genomes, and in diverse isolates within the dataset. No isolate harboured *ctxAB* in the absence of VPI-1, consistent with our understanding that VPI-1 encodes the TCP receptor to which CTXφ binds, notwithstanding transduction of CTXφ amongst TCP$^-$ *V. cholerae* [93]. T3SS elements were not mutually exclusive with either CTXφ or VPI-1. This is not the same as the observations made in Chapter 3, and underlines the fact that as more diverse *V. cholerae* are sequenced, our understanding of gene distribution within the species will be refined.

One clade of isolates was identified which harboured the T3SS most similar to a system from *V. anguillarum*, described in Chapter 3 (Figure 5.5). This cluster of isolates was polyphyletic, consisting of one clade of the three Argentinian genomes described in Chapter 3 and a second clade comprising an additional 11 non-O1 isolates from Mexico and Guatemala (eight of clinical origin, three of unknown origin). None of these genomes appear to contain any other known *V. cholerae* virulence determinants. This observation underlines further the fact that there is a great need to understand the contributions made by T3SS to disease caused by *V. cholerae* in humans – throughout this thesis research, several examples have been identified of clinically-isolated *V. cholerae* that lack putative virulence determinants other than T3SS, justifying further research into the roles that such systems may play in causing acute watery diarrhoea or symptoms of cholera.

5.3.4 – Serogroup assignment of isolates in silico

Complete metadata were unavailable for the sequenced isolates contained in this analysis. Therefore, the serogroup (O1 or non-O1) of each isolate was confirmed *in silico*. To do this, the presence and absence of each of the genes in the O1 biosynthesis gene loci was determined for each isolate in the dataset (Figure 5.6A). These loci have been delineated previously in *V. cholerae* (discussed in Chapter 4 in the context of *V. cholerae* O139). Isolates that are known not to be serogroup O1 – including serogroup O139 and O37 isolates – were confirmed to lack genes present in serogroup O1 isolates (Figure 5.6B). Based on these results, all of the isolates in the dataset were determined to be O1 or non-O1 *V. cholerae*.

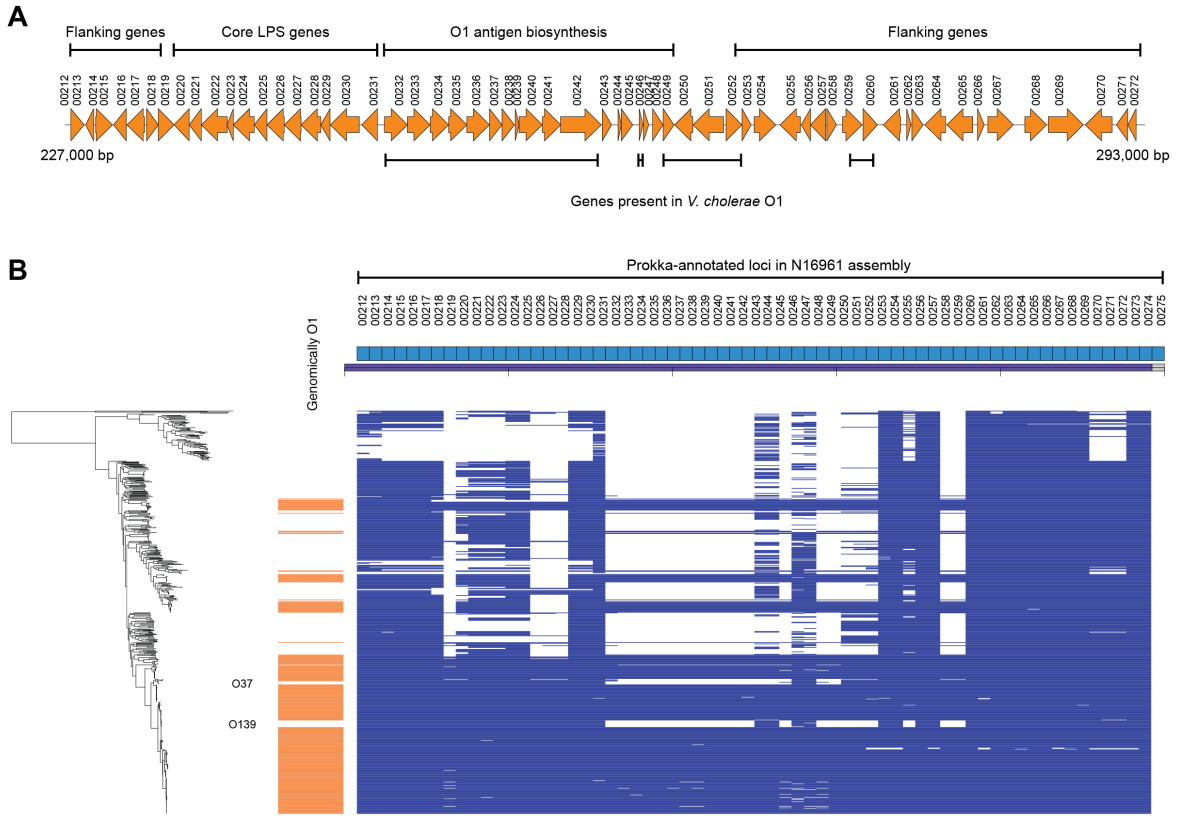


Figure 5.6 – The distribution of O1 antigen biosynthesis genes in the *V. cholerae* pangenome. (A): Schematic of the genomic organisation of the O1 biosynthetic locus in the N16961 reference sequence annotated with Prokka for uniformity within the pangenome dataset. The genes involved in producing the core lipopolysaccharide as well as the O1-specific LPS antigen are denoted. (B): The presence and absence of the 60 genes labelled in (A) across the pangenome of 651 isolates presented in Figure 5.1. Isolates which harbour all of the genes in (A) were determined in silico to be serogroup O1. Where required, this was also validated using BLASTn to search the assembly for the presence of the O1 biosynthesis locus in its entirety, as carried out in Chapter 3. This approach allows for the easy recognition of non-O1 *V. cholerae*; the O37 and epidemic O139 clusters (7PET sub-lineage) are indicated as examples.

5.3.5 – Distribution of key pathogenicity islands amongst V. cholerae

It has been stated repeatedly in the literature that the four major *V. cholerae* pathogenicity islands, VPI-1 and VPI-2, VSP-1 and VSP-2, are markers for epidemic and pandemic lineages of *V. cholerae* O1 [133, 142]. Having determined whether or not each isolate in the phylogeny was serogroup O1, and knowing from Chapter 4 that these elements may be present in more than one copy, the presence/absence and copy number of the genes encoded by these pathogenicity islands across the phylogeny was determined. This was done by testing for the presence of each of the genes that comprise these islands in the pangenome data, in order to obtain a sense of the variability within these elements across the phylogeny. These results are summarised in Figure 5.7.

Some previously-described phenomena were evident amongst these results, such as the VPI-2 deletion in *V. cholerae* O139 [134, 409], and the deletion in VSP-2 reported in recent 7PET isolates [309]. As previously observed, a number of non-7PET *V. cholerae* O1 from China were found to harbour VSP-1 in its entirety [396], as does a non-7PET *V. cholerae* O1 isolate from Thailand [440]. These results are of particular importance because they challenge the hypothesis that VSP-1 and VSP-2 are specific to 7PET [133]. It also appeared that genes orthologous to those encoded by VSP-2 were identifiable in multiple *V. cholerae* that were distantly related to 7PET (Figure 5.7). It was also evident that a number of non-O1 isolates in this dataset harboured VPI-1 (genes on which encode TCP), suggesting at least in principle that these non-O1 isolates could be lysogenised by CTXφ. However, no non-O1 isolates harbouring VPI-1 and CTXφ were detected amongst these data (Figure 5.7).

This analysis underscored the importance of using closed genome assemblies for such analyses. To compute this phylogeny, a short-read assembly for the toxigenic *V. cholerae* O139 described in Chapter 4 was used, rather than the corresponding long-read assembly. Consequently, the duplication of VSP-1 was not detected in this isolate, although its duplication had been demonstrated repeatedly in Chapter 4. However, VSP-1 was found to be duplicated in the closed genome of MJ-1236, consistent with a previous report of this isolate [98].

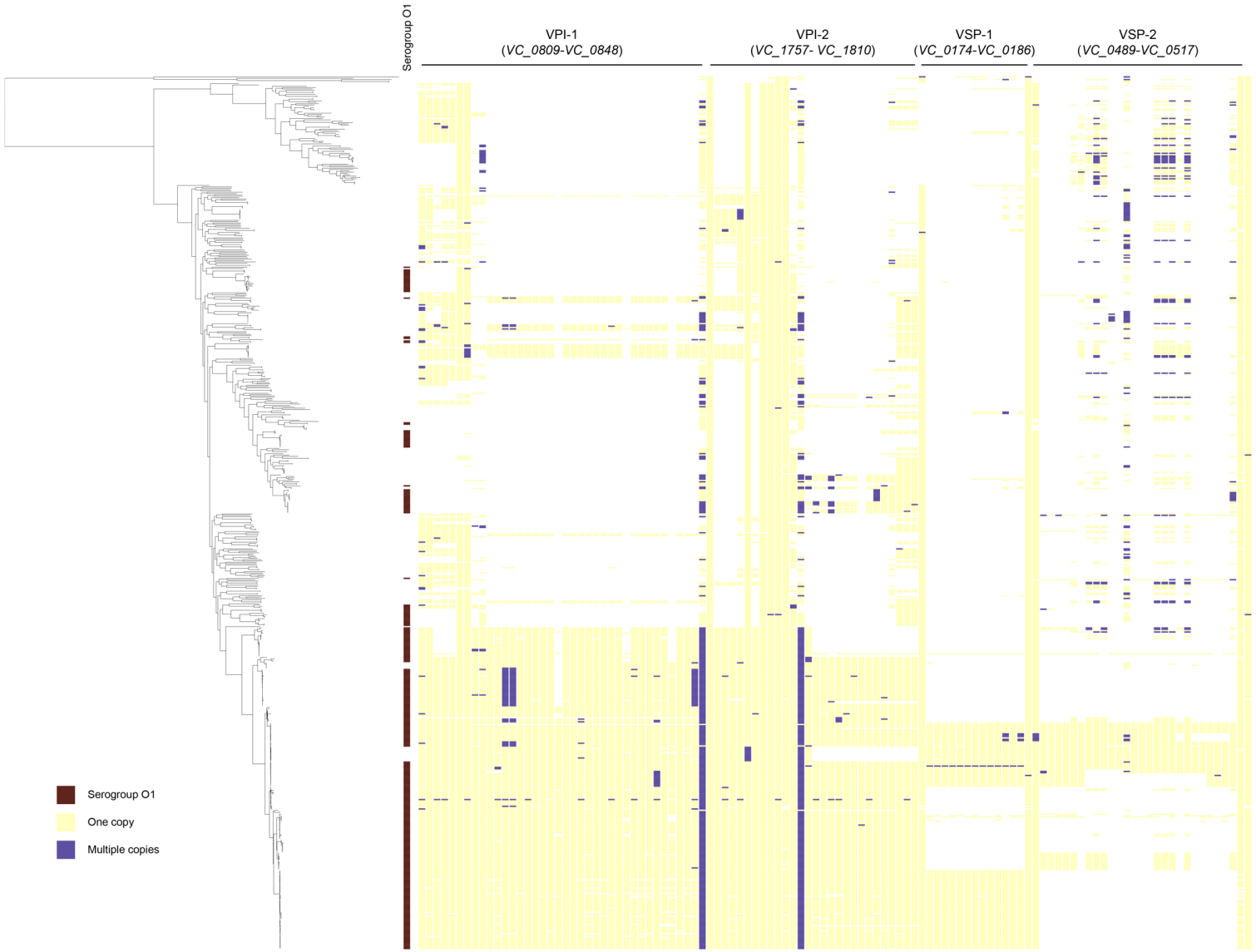


Figure 5.7 – Distribution of genes encoded by canonical pathogenicity islands in the *V. cholerae* pangenome. Legend on following page.

Figure 5.7 legend: The locus IDs from the N16961 reference genome for each pathogenicity island are indicated. Genes flanking these pathogenicity islands were also included, and are evident as genes at the beginning or end of each island which are present in nearly every isolate in the dataset. Serogroup assignment was taken from Figure 5.6. Since this plot reports the presence and absence of elements found in N16961, $tcpA^{El\ Tor}$ appears to be absent from genomes which harbour the $tcpA^{Classical}$ allele, though the remainder of VPI-1 genes are present in such genomes

5.3.6 – Plasmid and antimicrobial resistance gene distribution amongst V. cholerae

Continuing the exploration of these diverse genomes, all genome assemblies were scanned for the presence of antimicrobial resistance determinants and plasmid replicon sequences (see Methods). The results of this analysis are summarised in Figure 5.8. It was evident that isolates within the diverse clade (Figure 5.8, grey disc) harbour sequences homologous to the beta-lactamases bla_{CARB-7} and bla_{CARB-9}. These two genes have been discussed in Chapter 4; briefly, both are chromosomally-encoded within the integron on chromosome 2 [151, 152]. This is consistent with the lack of plasmid replicon sequences amongst these genomes (Figure 5.1, 5.8). In this analysis, NCTC 30 was determined to harbour bla_{CARB-7}, which corresponds to the functionally-validated $bla_{CARB-like}$ gene (section 4.3.9). Other observations that were consistent with previous reports include the presence of a *catB9* gene within 7PET [158], and the detection of *qnrVC* quinolone resistance genes amongst Chinese non-7PET genomes [396] (Figure 5.8).

It is important to note that the *tet(34)* sequence, predicted to be present in 650 of 651 isolates in this dataset, is homologous to a xanthine-guanine phosphoribosyltransferase (XPRT) gene present on the *V. cholerae* chromosome [59, 441]. This sequence was not detected in analyses for previous chapters, which made use of ARIBA to scan short-read data for resistance determinants (Chapter 3), and the Resfinder web interface, which scans assemblies using BLASTn (Chapter 4). Moreover, tetracycline resistance was not reported in the vast majority of phenotyped *V. cholerae* included in this dataset (Chapters 3, 4). Since this was the first time that ABRicate had been used to detect resistance genes in assemblies, it is possible that detection of *tet(34)* is an artefact of the chosen analysis method. Additional validation would be required before further inferences can be drawn from this observation.

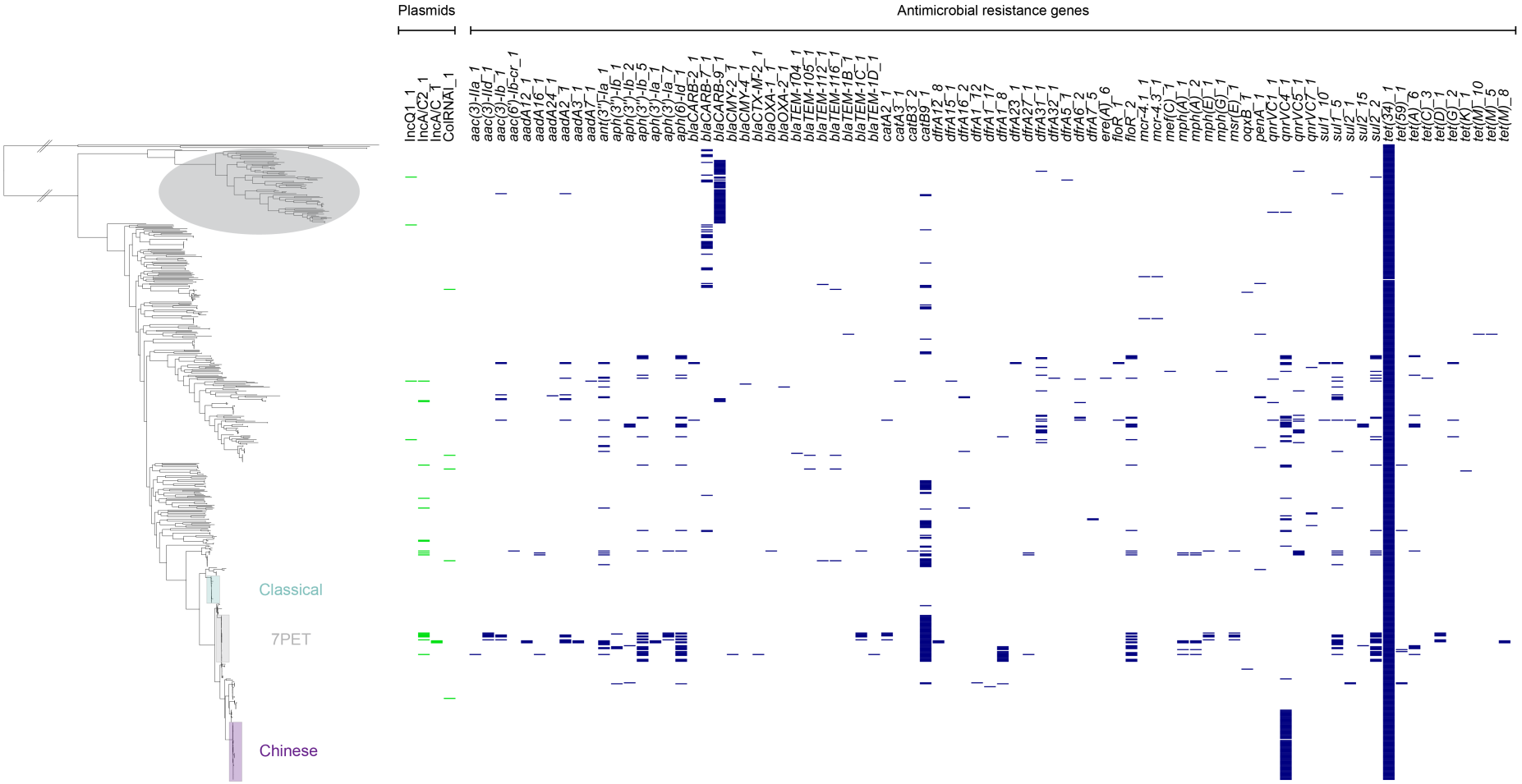


Figure 5.8 – Distribution of AMR genes and plasmid replicons within the *V. cholerae* phylogeny. Hatch marks denote branches which have been manually shortened for illustrative purposes. Assemblies were scanned for sequences of interest using ABRicate and the ResFinder/PlasmidFinder databases (see Methods). Plasmids, AMR genes >75% coverage and identity

From these data, it was evident that the majority of sequenced *V. cholerae* possessed very few antimicrobial resistance genes; 550 of the 651 genomes harboured two or fewer AMR determinants (Figure 5.9). The presence of a plasmid replicon is statistically significantly associated with a *V. cholerae* isolate harbouring three or more AMR genes (χ^2 (d.f. = 1, n = 650) = 75.84; $p < 0.00001$; Table 5.1). This suggests that the majority of AMR genes amongst this dataset are present on plasmids, of which there are very few types amongst these genetically heterogenous isolates.

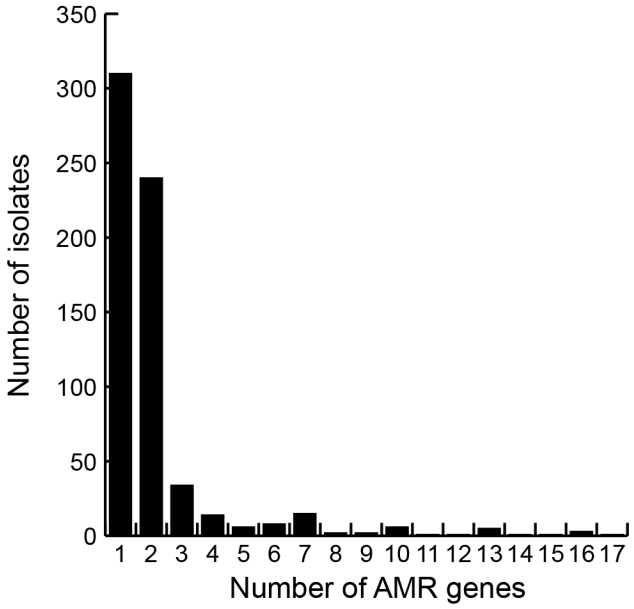


Figure 5.9 – The vast majority of *V. cholerae* isolates harbour two or fewer AMR genes. The potentially-dubious detection of *tet(34)* gene means that those isolates predicted to harbour one resistance gene are likely to be pan-susceptible.

	Plasmid Replicon		
AMR genes	No	Yes	Σ(rows)
≤ 2	542	8	550
≥ 3	79	21	100
Σ(columns)	621	29	650

Table 5.1 – χ^2 contingency table. Testing for association between the presence of a plasmid replicon (IncQ1, IncA/C1, IncA/C2, ColRNAI) and increased numbers of AMR genes.

It was evident that within this dataset were a number of serogroup O1 *V. cholerae*, many of which were part of a collection of *V. cholerae* supplied by NCTC. Many of these were the lineages local to Latin America that were previously-described [189], or were from a recent study of Chinese non-7PET *V. cholerae* O1 [396]. However, a number of non-7PET *V. cholerae* O1 were evident which were part of the NCTC collection of historical *V. cholerae* isolates. This prompted a closer investigation of the history and genomes of the NCTC *V. cholerae*, discussed below.

5.3.7 – Phylogenetic positions of historically-important NCTC isolates

As part of this thesis research, I spent time at PHE Colindale examining historical records for *V. cholerae* contained within the NCTC collection, and collating these data with the assistance of Jake Turnbull. Names and basic metadata for the 35 isolates included in this chapter are listed in Table 5.2. Following the detailed characterisation of the PacBio-assembly of NCTC 30 (Chapter 4), the phylogenetic position of all other *V. cholerae* that were in-stock and available from the NCTC collection was determined (Figure 5.1). All of these isolates were available as live cultures for experimentation under CL3 conditions, and were sequenced using both PacBio and Illumina technologies with gDNA prepared using the methodology optimised using NCTC 30 (Methods, section 2.2.5). All live isolates are denoted on Figure 5.1.

NCTC No.	Other strain references and names	Year of isolation	Internal isolate ID	Serogroup/Serotype from records	Phylogenetic lineage (Figure 5.1)	*In silico* serogroup
30	ATCC14735; MARTIN 1	1916	MJD382	Non-O1/O139 (O2)	-	Non-O1
3661	CN 5870; DOORENBOS 80	1931	MJD383	O1 El Tor	near-ELA5	O1
4693	JAPANESE ORIGINAL	pre-1936	MJD384	O1 Inaba	Classical	O1
4711	ATCC 14730; NANKING 32/123	Pre-1936	MJD385	O2	-	Non-O1/O139
4714	EL TOR 34-D 19	1934	MJD386	Non-O1/O139	-	Non-O1/O139
4715	ATCC 14731; CN 3426; ELTOR 34-D 23	1934	MJD387	O3	-	Non-O1/O139
4716	ATCC 14732; KASAULI 73	1932	MJD388	O4	-	Non-O1/O139
5395	ATCC 14734; IRAQ	1938	MJD389	O1 El Tor Ogawa	pre-7PET	O1
5596	SHANGHAI 10	pre-1939	MJD390	O1 Ogawa	Classical	O1
6585	SUBAMMA	pre-1944	MJD391	No data	Classical	O1
7258	EGYPT 109	pre-1948	MJD392	No data	Classical	O1
7260	EGYPT 117	pre-1948	MJD393	No data	Classical	O1
7270	HIKOJIMA	pre-1948	MJD394	O1 Inaba	Classical	O1
8021	ATCC 14035; RH 1094	pre-1950	MJD395	O1 Classical Ogawa	Classical	O1
8039	CAIRO 1A	pre-1950	MJD396	O1 Classical Inaba	Classical	O1
8040	726/575 A	pre-1950	MJD397	O1 Classical Inaba	Classical	O1
8041	757 AUTOAGGLUTINABLE	pre-1950	MJD398	O Rough	Classical	O1
8042	ATCC 14733; WDCM 00203	pre-1950	MJD399	Non-O1/O139	-	Non-O1/O139
8050	CALCUTTA	pre-1950	MJD400	No data	-	O1
8367	4 Z	pre-1952	MJD401	O1 Classical Ogawa	Classical	O1
8457	ATCC14033; CN5774; DO1930; RH1092	1930	MJD402	O1 El Tor	near-ELA5	O1
9420	CN 5789; TOR A	pre-1955	MJD403	El Tor	pre-7PET	O1
9421	CN 5790; TOR 1	pre-1955	MJD404	O1 El Tor Ogawa	pre-7PET	O1
9422	CN 5871; TOR 8	pre-1955	MJD405	O1 Inaba	near-ELA5	O1
9423	TOR 31	pre-1955	MJD406	No data	pre-7PET	O1
10255	CN 5745; 20109	1961	MJD365	O1 El Tor Ogawa	7PET	O1
10256	CN 5748; 20111	1961	MJD366	O1 El Tor Ogawa	7PET	O1
10732	CN 3534; 384/52	1952	MJD367	O1 Classical Inaba	Classical	O1
10733	AJ 1592; CN 3539; I-5; 586/52	1952	MJD368	O1 Classical Ogawa	Classical	O1
10954	1330	1973	MJD369	El Tor	7PET	O1
11090	NCIB 9341; 1077	pre-1950	MJD370	No data	-	O1
11348	VL3029; WDCM00136, CCUG67718, DSM101014	pre-1981	MJD372	O24	-	Non-O1/O139
11500	VL 7050	pre-1983	MJD373	No data	-	Non-O1/O139
11643	VL 4944	pre-1985	MJD380	No data	-	Non-O1/O139
12946	ATCC 51395; MO3	1993	MJD381	O139	7PET (O139)	O139

Table 5.2 – NCTC *V. cholerae* sequenced for this thesis research. Excluded from this list is one sequenced isolate which was found to be a member of the *Aeromonas* genus upon analysis of the genome assembly.

5.3.7.1 – Pandemic NCTC isolates

Of the 35 NCTC isolates that were included in this analysis, four were found to be members of 7PET (Figure 5.10). One of these was NCTC 12946, also known as MO3, a serogroup O139 strain isolated at the beginning of the *V. cholerae* O139 epidemic in 1993 [436]. Genomic analysis confirmed that this isolate possessed serogroup O139 LPS genes, was toxigenic, and was phylogenetically positioned amongst previously-sequenced toxigenic *V. cholerae* O139 within 7PET (Figure 5.10). NCTC 10255 and 10256 were isolated from patients during an outbreak of "paracholera" in Hong Kong during 1961 [358]. Vella demonstrated that these El Tor Vibrios were highly virulent in a mouse model, and that these isolates showed some ability to immunise mice against re-infection [358]. NCTC 10954 has been used historically as a positive control strain for haemagglutination in *V. cholerae* biotyping [40].

Thirteen isolates were members of the Classical lineage (Figure 5.10). Vella compared NCTC 10255 and 10256 to a "true cholera" vibrio, NCTC 7260, the most virulent isolate available at the time [358] (Table 5.1; Figure 5.10). The finding that this is a Classical isolate is logical, given that Vella had referred to this as a "true cholera" *Vibrio* when compared to El Tor isolates [358]. Other isolates of interest that were members of the Classical lineage include NCTC 8367 (strain 4Z), which has been used as a source of neuraminidase ("receptor destroying enzyme") [135, 442]. Neuraminidase is encoded by *nanH*, part of VPI-2 [134] (section 1.2.5), a pathogenicity island found in Classical *V. cholerae* [54, 133, 134] and in this isolate (Figure 5.7). Characterising the phylogenetic position of NCTC 8021 is particularly important because this isolate is the *Vibrio cholerae* type strain and the neotype of the *V. cholerae* species [443]. In 1965, Hugh compared the biochemical phenotypes of 258 *V. cholerae* isolates to that of NCTC 8021 [444], and requested an opinion from the Judicial Commission of the International Committee on Bacteriological Nomenclature that this strain be considered the *V. cholerae* neotype [209]. *V. cholerae* appeared on the Approved List of Bacterial Names in 1980 [35], and Hugh's work was cited as the description of the *V. cholerae* species [35, 444]. NCTC 8021 has been used as a reference standard for the negative haemagglutination phenotype used to biotype Classical *V. cholerae* [40]. It is reassuring that this type strain is a member of the Classical lineage, given that NCTC 8021 was designated a type strain during the 1960s, at a time when Classical *V. cholerae* was still considered to be the sole aetiological agent of cholera [1, 36]. Thus, the type strain chosen was indeed biochemically, microbiologically, and genomically, a representative of the aetiological agent of Asiatic cholera.

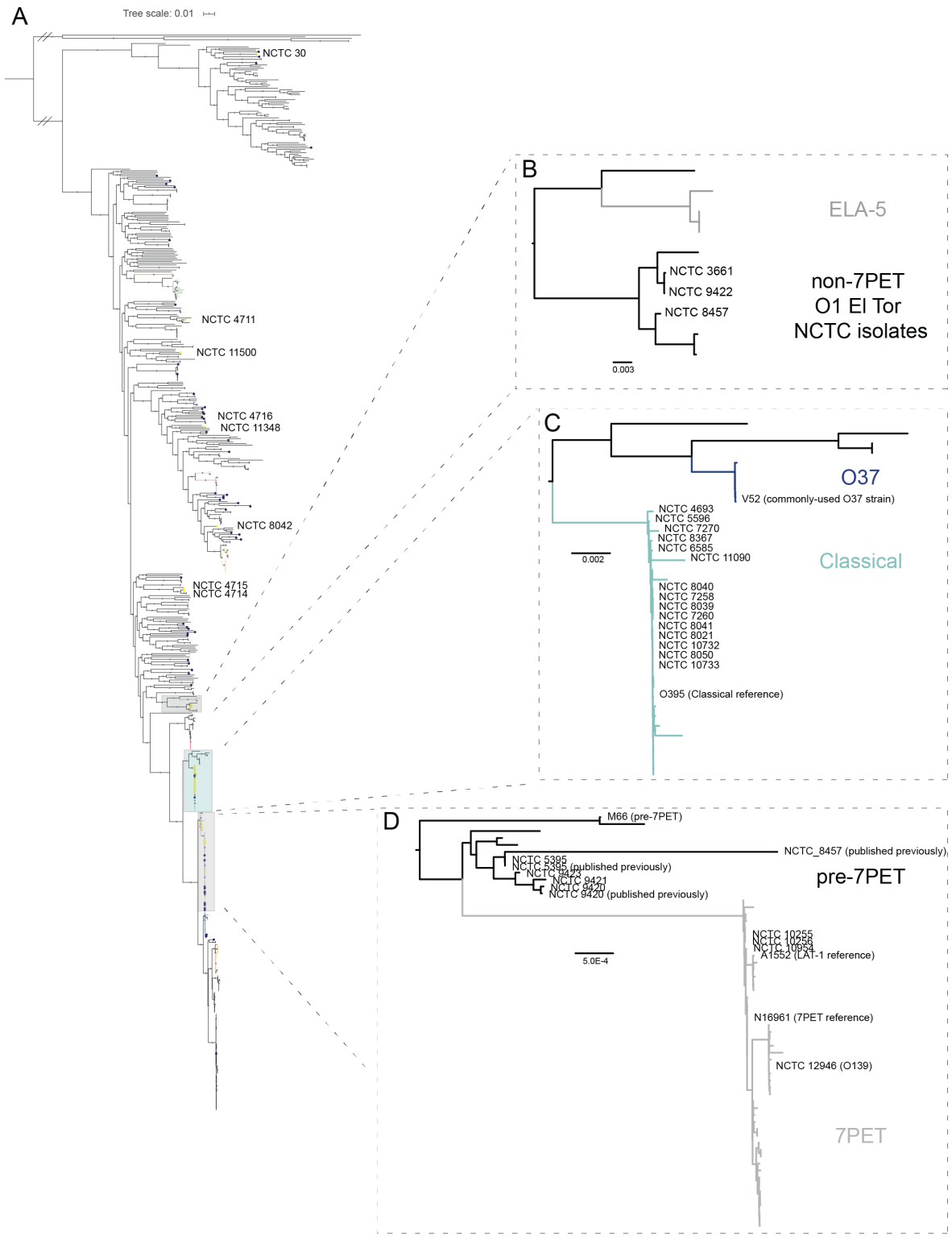


Figure 5.10 – A *V. cholerae* phylogenetic tree annotated with the names and IDs of NCTC isolates. The phylogeny presented in Figure 5.1 is re-drawn and the position of NCTC isolates indicated. Where necessary for illustrative clarity, sub-trees were extracted from the original phylogeny and visualised separately (Figtree v1.4.3). The names of NCTC isolates and key reference sequences have been retained. Hatch marks indicate branches that were manually shortened for illustrative purposes.

5.3.7.2 – Non-pandemic NCTC V. cholerae *O1*

Of interest was a set of three *V. cholerae* O1 which were not part of either the 7PET or Classical lineages. NCTC 3661, 8457, and 9422 were all recorded as being serogroup O1 in NCTC records (Table 5.2). NCTC 3661 and 8457 were also recorded as being biotype El Tor (Table 5.2). Genomic analysis confirmed that these three isolates harboured the genes necessary to make these serogroup O1, but lacked VPI and VSP pathogenicity islands and CTXφ, though they were predicted to harbour T3SS-2α (Figure 5.6, 5.7, 5.8). In the absence of any further clinical metadata, it is impossible to know whether the T3SS-2α predicted to be present in these isolates is either functional or contributed to causing disease in a patient.

These three isolates were particularly interesting because of their history. NCTC 3661, Doorenbos 80, was isolated in 1931 from a healthy Mecca pilgrim [41]. Gardner and Ventakraman described this as having 'typical' biochemical characteristics, being of O-subrgoup I, and to be capable of haemolysing goat erythrocytes [41]. In subsequent studies, NCTC 3661 was also shown to be haemolytic on sheep blood agar, produced a positive Voges-Proskauer test result, and was positive in a Grieg test for haemolysis [445]. However, this strain has been shown to be sensitive to group IV cholera bacteriophage, a phage to which El Tor isolates tend to be resistant [446].

NCTC received a batch deposition of four *V. cholerae* from A.H. Wahba, which were received simultaneously from Cairo in 1953 (NCTC records). At the time, these isolates were named Tor A, Tor 1, Tor 8 and Tor 31 (Table 5.2), and were simply described as "El Tor strains" (NCTC records). These were accessioned into the NCTC collection as NCTC 9420-9423. A recent study reported genome sequences for NCTC 9420 and NCTC 5395, and found that both of these sequences were basal to the 7PET lineage [213], dubbed "pre-7PET" here. These sequences are included in the phylogeny presented in Figure 5.10. It has been stated that the properties of NCTC 9420, as well as NCTC 8457 and 5395, are likely to resemble those of the original El Tor biotype *V. cholerae* described by Gotschlich, which tended to be isolated from asymptomatic patients [211, 213]. For this PhD project, NCTC 9420, 9421 and 9422 were sequenced. Although NCTC 9420 and 9421 are basal to 7PET, isolate NCTC 9422 sits adjacent to NCTC 3661 (Figure 5.10). This strongly suggests that the El Tor phenotypes, which led NCTC 9420-9423 to be isolated and deposited with NCTC, are consistent between the pre-7PET isolates and the phylogenetically-unrelated NCTC 9422.

A closer inspection of the NCTC 3661 and NCTC 9422 genome sequences demonstrated that both isolates were predicted to harbour IncA/C2 plasmids (Figures 5.1, 5.8). This was surprising, because when the assemblies for both of these isolates were scanned for antimicrobial resistance determinants, no putative resistance genes were detected (Figure 5.8). To address this, a combination of PacBio RSII long-reads and Illumina short-reads were used to produce hybrid assemblies for these genomes. Unicycler, with 'conservative' settings, was used to minimise the likelihood that contigs would be merged spuriously (see Methods). A ~130 kb circularised contig was identified in the hybrid assemblies of both genomes. Visualisation of the De Brujin graph from the corrected, circularised assembly confirmed that this element was independent of the two *V. cholerae* chromosomes (see Figure 5.11 for representative result from NCTC 3661), and this contig contained the IncA/C2 replicon. No antimicrobial resistance genes were detected on this contig.

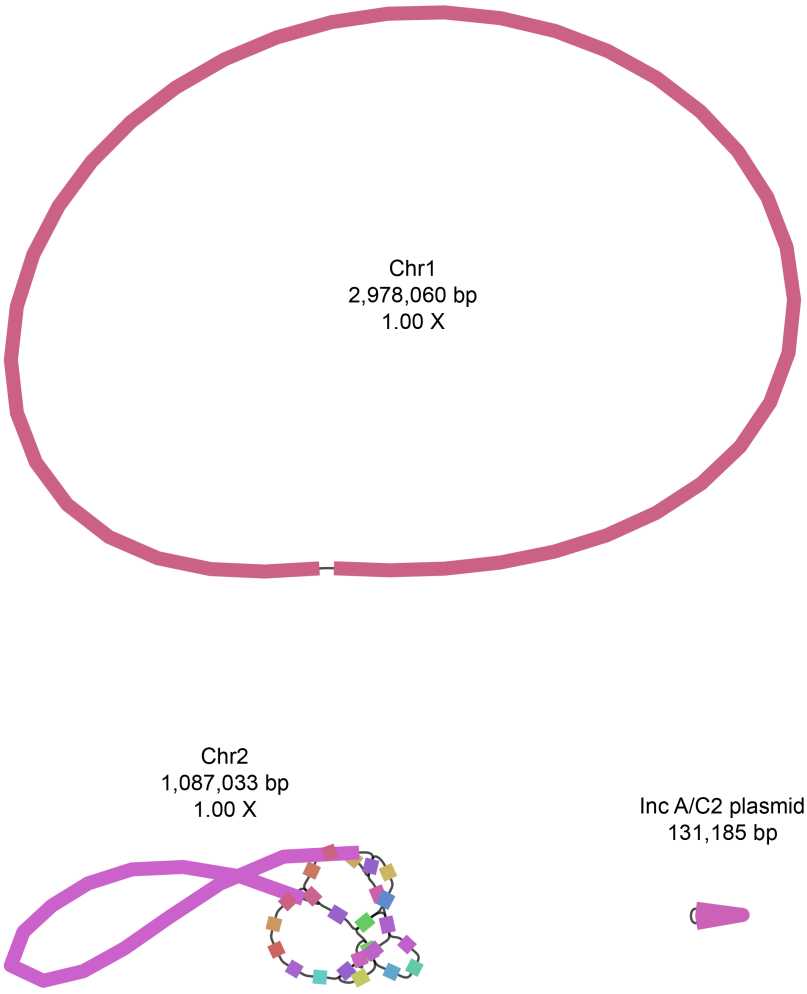


Figure 5.11 – Visualisation of the De Brujin graph for the polished, rotated hybrid assembly for NCTC 3661. The integron on chromosome 2 was not completely assembled; this is a consequence of running Unicycler on 'conservative' settings. An identical plasmid of 131,185 bp was assembled from NCTC 9422.

Comparative genomics demonstrated the similarity between this putative IncA/C2 plasmid (pNCTC3661) and previously-published *V. cholerae* plasmids from the same family, including the rare MDR IncA/C2 plasmid detected in Argentinian LAT-1 *V. cholerae* described in Chapter 3 (Figure 5.12). It also highlighted that the variable regions of the plasmid backbone, which can harbour antimicrobial resistance genes, are much shorter in this drug-sensitive plasmid compared to plasmids encoding resistance determinants (Figure 5.12).

5.3.7.1 – NCTC 8457

The first report of NCTC 8457 was made by Doorenbos and Kop, in 1951, in which paper the isolate is recorded as having been isolated 20 years prior to the publication [447]. This strain is serogroup O1, biotype El Tor, and non-toxigenic [213]. On this basis, Hugh argued that NCTC 8457 should be the neotype of *Vibrio* eltor Pribram (*i.e.*, El Tor *V. cholerae*), since this isolate displayed all of the characteristics of an El Tor vibrio [210]. A genome sequence for NCTC 8457 was reported in 2009 [54] but was first sequenced in 2006 at TIGR (accession # NZ_AAWD00000000.1). This sequence has subsequently been used as an exemplar of the "pre-seventh pandemic" group of *V. cholerae* [54, 213]. In particular, this strain was used alongside other historical isolates to make inferences about the evolution of the lineage now dubbed 7PET [213].

Unlike our sequence data for NCTC 9420 and 5395, the phylogenetic position of our sequenced isolate of NCTC 8457 was not consistent with that of the previously-reported genome (Figure 5.10). Rather than co-clustering with the previously-sequenced NCTC 8457 isolate as did our re-sequenced NCTC 5395 and 9420 (Figure 5.10D), our stock of NCTC 8457 clustered with the two serogroup O1 biotype El Tor isolates NCTC 9422 and 3661 (Figure 5.10B). This was a surprising result, but was replicated when gDNA from an independent batch of NCTC 8457 (DNA supplied by M.A. Fazal at NCTC) was sequenced and similarly analysed. This suggested that there may be discrepancies between the previously-sequenced culture of NCTC 8457 and the stocks which were sequenced for this PhD project.

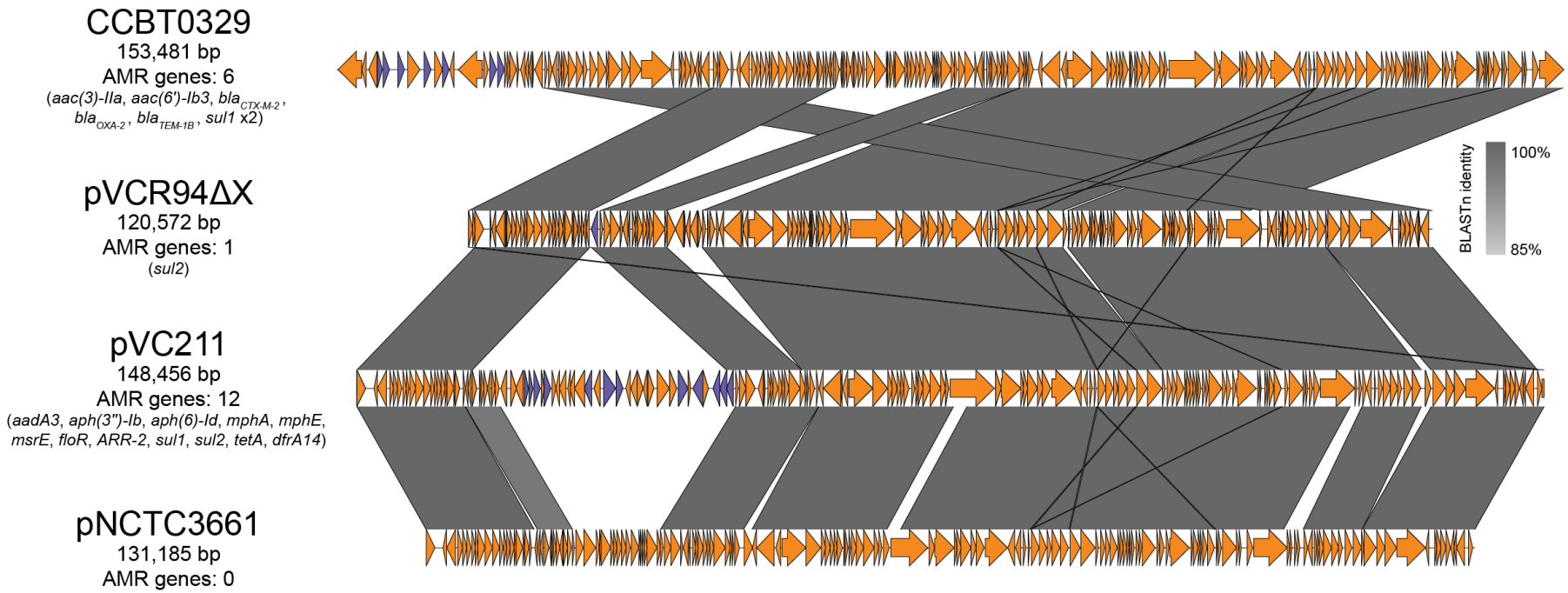


Figure 5.12 – Comparing pNCTC3661 to three *V. cholerae* IncA/C2 plasmids. Included in this comparison is the Argentinian IncA/C2 plasmid described in Chapter 3 (Figure 3.20).

In light of this apparent discrepancy, a number of confirmatory analyses were performed to validate our genomic observations, and to provide support for the phylogenetic position of this sequence. Firstly, it was confirmed *in silico* that NCTC 8457 was likely to be of serogroup O1 (Figure 5.6). This is fully-consistent with previous reports [54]. Similarly, a manual inspection of the genome confirmed that CTXφ and the genes encoding cholera toxin could not be detected in this genome (Figure 5.7); again, consistent with previous characterisations of the isolate.

Interestingly, a conservatively-generated hybrid assembly for NCTC 8457 suggested the presence of a small circular element in this assembly, which was independent of the two chromosomes (Figure 5.13). No plasmids were detected in the hybrid assembly using the Plasmidfinder website or in the earlier analysis of the short-read assembly (Figure 5.8).

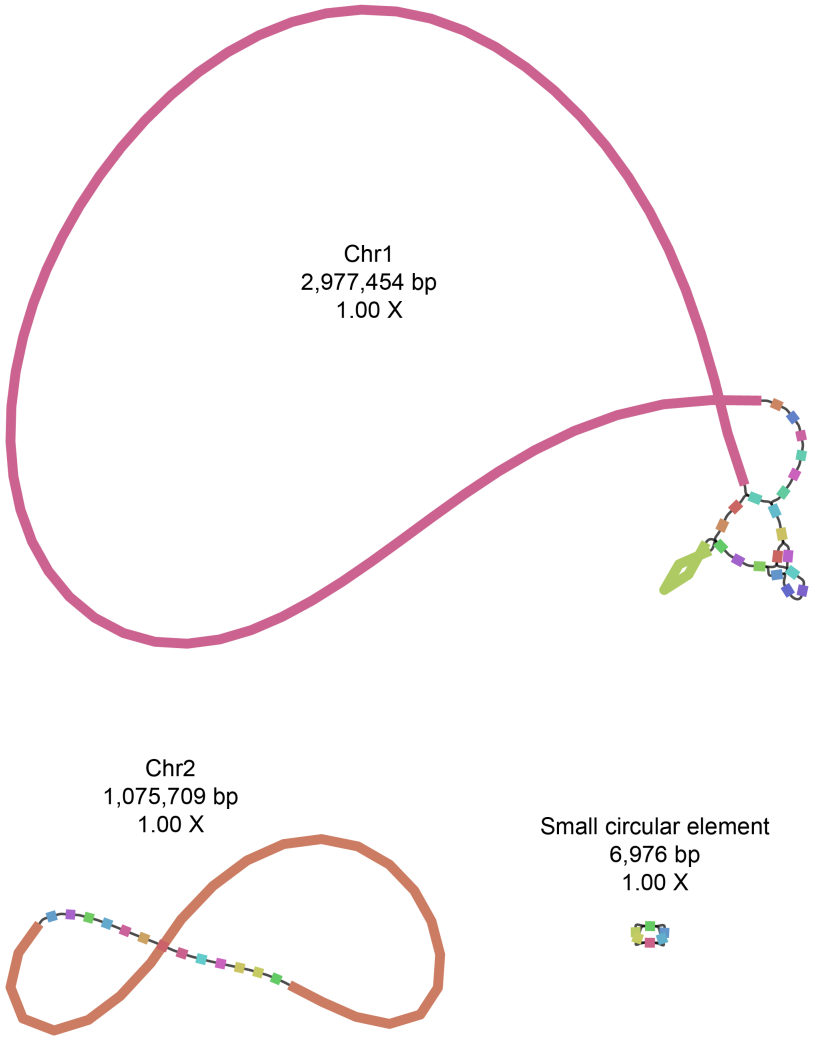


Figure 5.13 – Visualisation of the De Brujin graph for the NCTC 8457 hybrid assembly. Figure produced using Bandage and the polished assembly graph generated by Unicycler.

The six contigs that comprised this 6.9 kb element were extracted from the hybrid assembly, and their concatenated sequence was used to query the nr database using BLASTn. The most similar sequence to this element was a 3.8 kb plasmid called pVC (accession # AY423429), the sequence of which covered the query from NCTC 8457 by 69%, with 98.99% nucleotide identity. pVC is a cryptic plasmid which was isolated from an environmentally-derived non-O1/O139 *V. cholerae* isolate dubbed MP-1 [448]. This manuscript reported that four CDSs of unknown function were identifiable on pVC, that attempts to cure pVC from MP-1 were unsuccessful, and that pVC was likely to replicate *via* theta replication [448, 449]. The replication origin of pVC was identified in this work [448], and this sequence is shared between pVC and the putative plasmid from NCTC 8457.

In order to confirm further whether the 6.9 kb element in NCTC 8457 might be an extrachromosomal plasmid, plasmid extractions were performed using stationary-phase cultures of this strain and others (alkaline lysis using Qiagen Midiprep kits; Methods, section 2.2.14). A small DNA element was successfully extracted from cultures of NCTC 8457 (stock ID MJD402; Figure 5.14), further suggesting that this isolate harbours a small plasmid that replicates extrachromosomally. Additional support for these observations was obtained while reviewing the literature surrounding NCTC 8457 - a previous publication had reported that NCTC 8457 harboured a small cryptic plasmid [450]. In future work, excising this DNA element and using a combination of Sanger sequencing and primer walking may allow for the identity and sequence of this element to be confirmed.

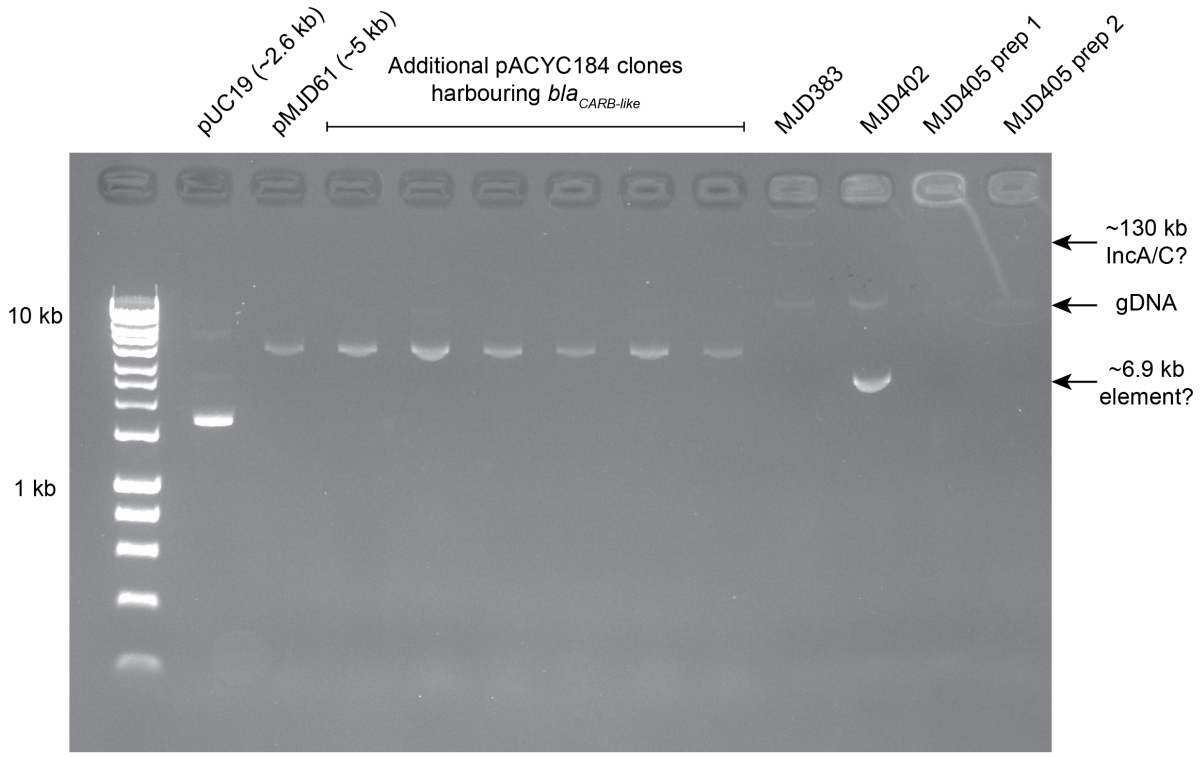


Figure 5.14 – Gel of plasmid preps from *V. cholerae* and *E. coli*. The plasmid preparations presented here were made from different organisms (*E. coli* and *V. cholerae*). Plasmids of known size from *E. coli* (lanes 1-8) were extracted from 2 ml stationary-phase overnight cultures using a Qiagen miniprep kit (Methods). Plasmids were extracted from 25 ml cultures of *V. cholerae* (lanes 9-12) using a Qiagen midiprep kit (Methods). Plasmids were not linearised before electrophoresis and host organisms were not consistent; hence, sizes are approximate. Ladder: Hyperladder 1 kb.

5.3.8 – Biotype determinants

The collection of historically-important isolates described above led us to investigate the basis of biotyping across this phylogeny. We chose to do this because (a) a number of Classical genomes had been added to the collection, and (b) three O1 El Tor isolates had been characterised which were distantly related to 7PET. Understanding the distribution of biotype-determining mutations across the species is an important step towards understanding what the "default" biotype would be for a newly-emerging O1 lineage, though it is accepted that biotyping reactions are uninformative when used on non-O1 *V. cholerae* [206]. The three principal phenotypes used to biotype *V. cholerae* O1 are sensitivity to polymyxin B, the Voges-Proskauer test, and haemolysis [40, 206]. The molecular and genetic bases of each of these phenotypes have been characterised, and were investigated across the phylogeny.

5.3.8.1 – Voges-Proskauer test and acetoin biosynthesis in V. cholerae

The Voges-Proskauer test detects acetoin in bacterial cultures [451]. Acetoin is produced as an intermediate molecule by *V. cholerae* during the fermentation of glucose to 2,3-butanediol [452]. El Tor *V. cholerae* produce acetoin and test positive when subjected to the Voges-Proskauer test, and classical *V. cholerae* do not produce acetoin, producing negative Voges-Proskauer test results [206]. The inability of classical *V. cholerae* to ferment glucose (*via* acetoin) means that classical biotype strains exhibit a severe growth defect when cultured on minimal media containing glucose as a sole carbon source, acidifying their growth media [452].

It has been shown that the AphA transcription factor negatively regulates acetoin production in *V. cholerae* [453]. This occurs because AphA directly repressed the *alsD* gene and the gene encoding the AlsR transcriptional regulator [453]. Additionally, the ability to produce (p)ppGpp has been shown to regulate acetoin production; this is independent of AphA [454]. It has also been shown that HapR, the master regulator of *V. cholerae* quorum-sensing responses, represses *aphA* when cultures reach a high cell density [455]. Thus, at high cell density, HapR represses *aphA*, thereby de-repressing the genes whose products are required to ferment acetoin (Figure 5.15) [453]. It has been shown that a single G→T mutation at position -77 in the *aphA* promoter ablates the ability of HapR to bind to, and repress, transcription of *aphA* [455]. It has been demonstrated that the -77T mutation in P_{aphA} is present in the Classical *V. cholerae* reference strain, O395, and that the -77G allele is present in 7PET isolate C6706 [455].

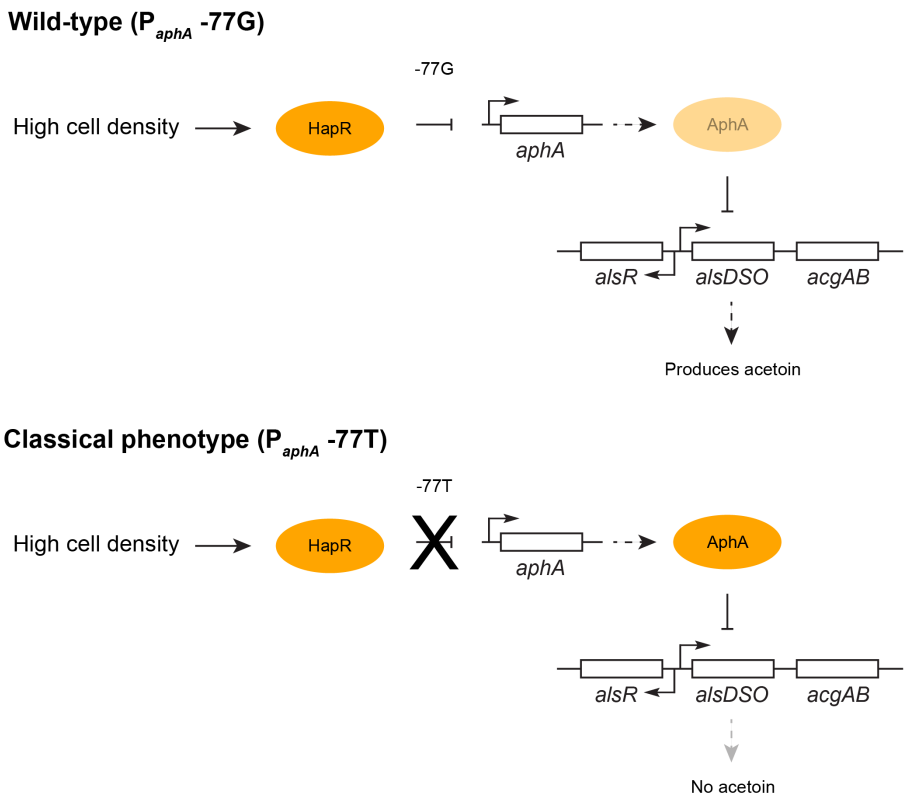


Figure 5.15 – Overview of the effects of P_{aphA} alleles on the expression of acetoin metabolism genes. At high cell density, HapR levels accumulate, and repress transcription of *aphA* if HapR binds to P_{aphA} (only if -77G allele is present). Diminished rates of *aphA* transcription and AphA translation leads to a concomitant de-repression of the *als* genes required to ferment acetoin.

In order to determine the P_{aphA} variants present in each isolate included in this dataset, *in silico* PCR was used to extract the *aphA* promoter sequence from each genome assembly (parameters listed in Methods, section 2.1.17). P_{aphA} was successfully extracted from 632 of the 651 sequences in the dataset. These sequences were aligned and compared to identify the variable sites in the promoter (Figure 5.16).

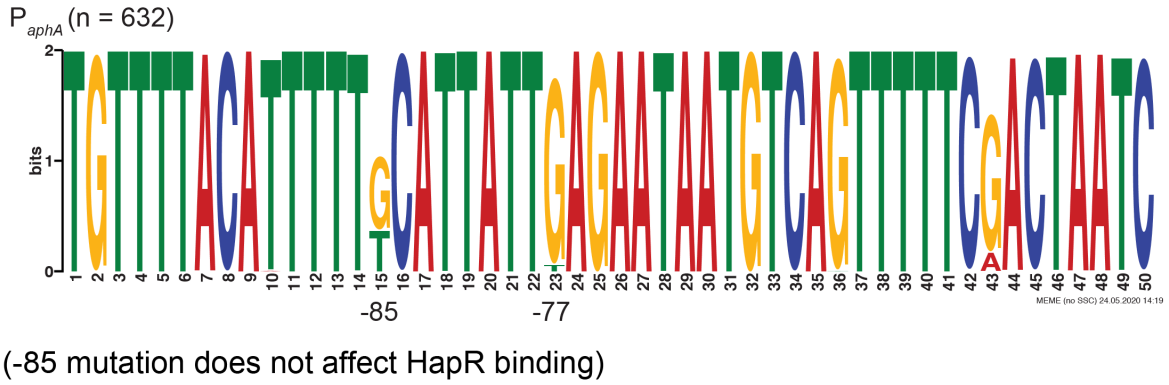


Figure 5.16 – P_{aphA} motif generated from a Clustal Omega alignment of 632 P_{aphA} sequences extracted using *in silico* PCR. The 50 bp region of interest was extracted from the "amplified" sequence using trimal. The MEME web server was used to produce the motif figure.

Two variable sites were visible in the P_{aphA} sequence that had been previously characterised – these were position -85, known not to influence HapR binding [455], and position -77, mutation of which from G to T abolishes HapR binding to P_{aphA} [455].

5.3.8.2 – Cholera toxin expression: an additional contrasting phenotype between classical and El Tor biotype isolates

The original study of *aphA* regulation by HapR was carried out in order to understand how HapR contributed to virulence gene expression in *V. cholerae* through AphA, and how this varied in a biotype-specific manner [455]. Canonically, classical biotype *V. cholerae* express the *toxR* virulence regulon *in vitro* at much higher levels than do El Tor isolates, which need to be cultured under stringent conditions to induce expression of virulence genes and CT *in vitro* [456, 457]. At low cell density, AphA and AphB co-ordinately activate the *tcpPH* promoter, and the binding of AphA to P_{tcpPH} enhances the binding of AphB to P_{tcpPH} [458, 459]. Resultant increases in TcpP and TcpH levels lead to an upregulation of *toxT* and the activation of virulence gene expression [458]. In classical isolates, the -77T mutation in P_{aphA} prevents *aphA* from being repressed by HapR; therefore, in classical isolates, the virulence regulon is not repressed at high cell density. In El Tor isolates, P_{aphA} is sensitive to HapR repression; therefore, at high cell density, AphA-mediated activation of P_{tcpPH} is reduced [458].

The differential activation of P_{tcpPH} by AphB is known to be necessary and sufficient to drive the classical and El Tor differences in virulence gene expression [460]. This has been shown

to be due to a single mutation in P_{tcpPH} ; the binding affinity of AphB for this promoter is tenfold higher if an A is present at position -65 [459]. In the O395 classical isolate, an A is present at -65, and in C6706, a G is present at the equivalent position (-66). Mutagenesis of both genetic backgrounds has shown that these alleles are sufficient to confer an El Tor or a classical virulence gene expression phenotype on *V. cholerae* [459]. This is summarised in Figure 5.17.

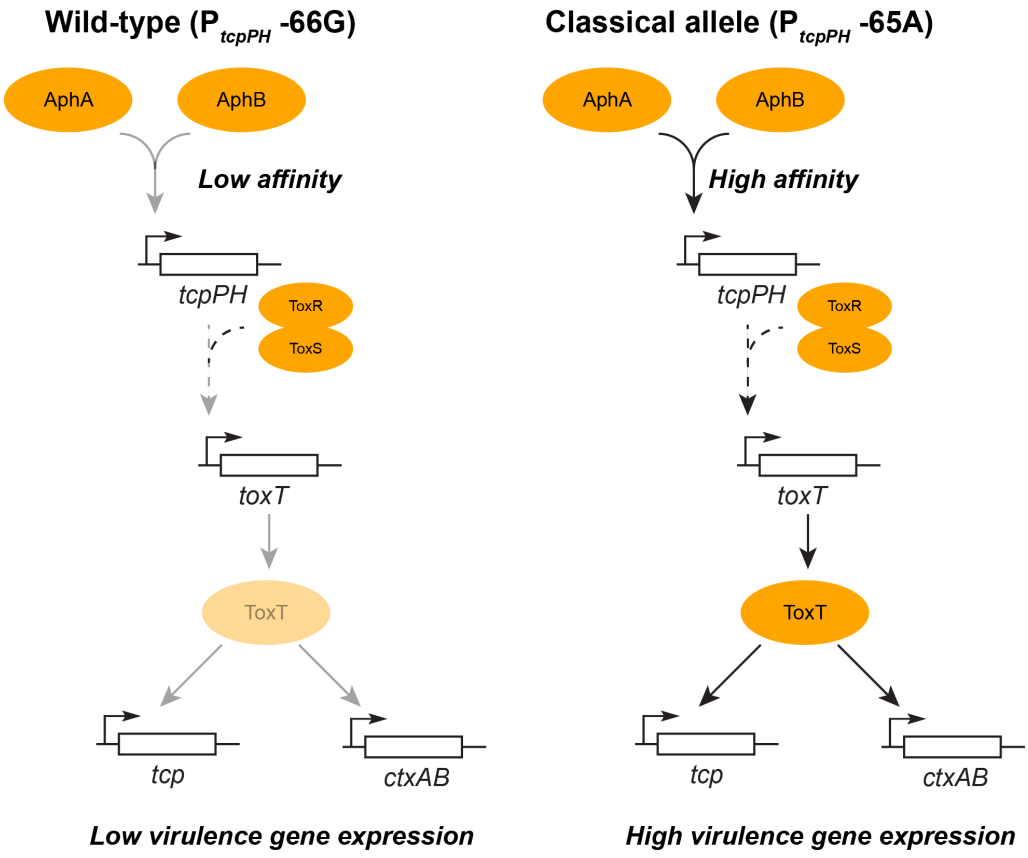


Figure 5.17 – Overview of the effects of P_{tcpPH} alleles on virulence gene expression. AphB (using AphA as a binding partner) has greatly enhanced affinity for P_{tcpPH} if an A is present at position -65/66. AphA/AphB binding to P_{tcpPH} activates expression of *tcpPH* and elevates intracellular levels of TcpP and TcpH. These, together with ToxRS, activate expression of *toxT*, thereby directly activating the expression of virulence genes. Conversely, in an "El Tor" background (P_{tcpPH} -66G), AphA and AphB have a lower affinity for P_{tcpPH}; therefore, virulence gene expression is lower.

It is important to consider mutations in both P_{tcpPH} and P_{aphA} at once, because together they couple the expression of virulence genes to quorum-sensing (Figure 5.18). In a Classical genetic background such as that of O395, *aphA* is rendered insensitive to repression by HapR. Therefore, regardless of cell density, *aphA* will be transcribed and AphA will be translated. Combined with the increased affinity for P_{tcpPH} experienced by AphA and AphB in this

background, this will lead to maximal expression of the *toxR*/*toxT* virulence regulon (Figure 5.18).

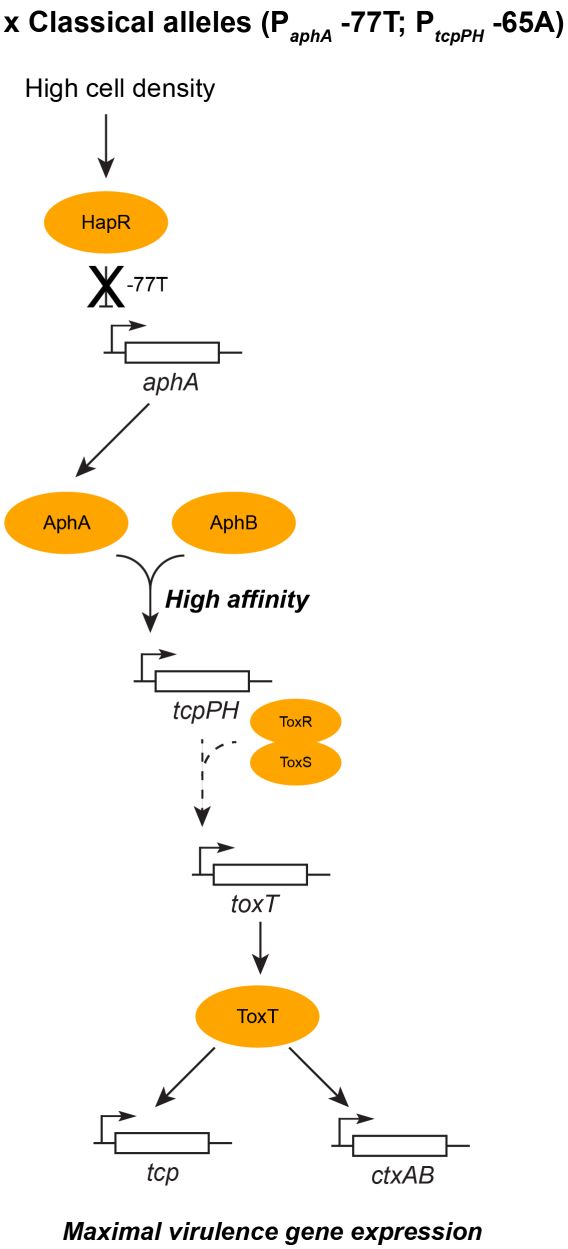


Figure 5.18 – Coordinate effects of Classical P_{tcpPH} and P_{aphA} alleles on virulence gene expression. The presence of both promoter alleles is predicted to lead to very high expression of the ToxT regulon.

In a similar manner to the study of P_{aphA}, the P_{tcpPH} sequence was extracted from all genomes in the dataset that harboured VPI-1 (258 isolates total), and compared to identify the variable sites in this promoter. This is illustrated in Figure 5.19.

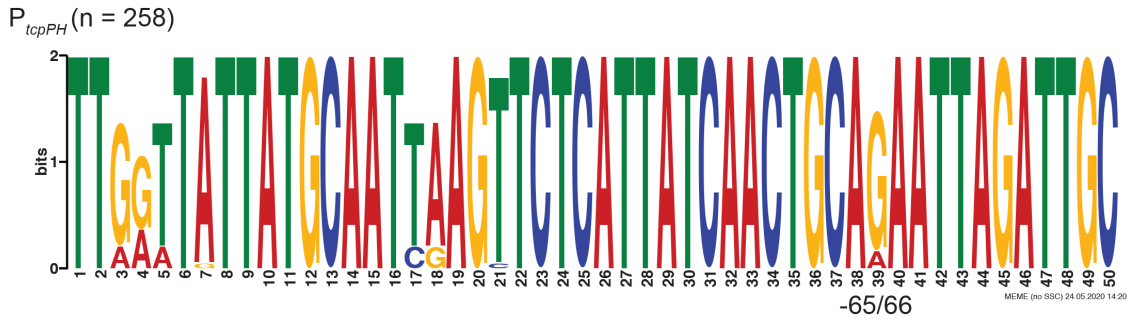


Figure 5.19 – P_{tcpPH} motif generated from a Clustal Omega alignment of 258 P_{tcpPH} sequences extracted using *in silico* PCR. The 50 bp region of interest was extracted from the "amplified" sequence using trimal. The MEME web server was used to produce the motif figure.

The genotype of P_{aphA} and P_{tcpPH} was determined for each genome in the dataset and mapped against the phylogeny, to determine the distribution of these variants (Figure 5.20). It was immediately apparent that both P_{aphA} -77T and P_{tcpPH} -65A were exclusively detected in Classical lineage *V. cholerae*, whereas the P_{aphA} -85 position was variable across the phylogeny.

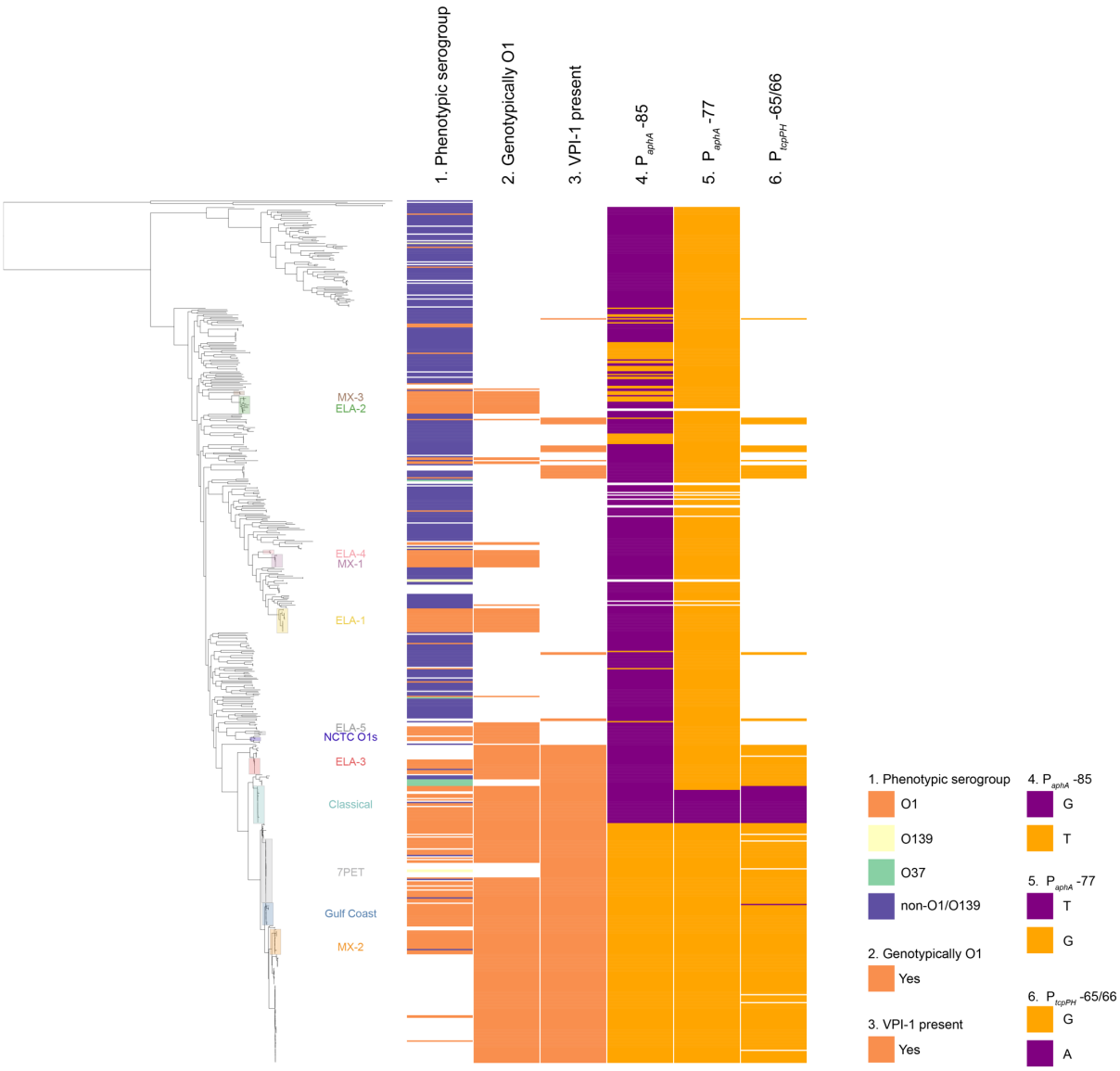


Figure 5.20 – Distribution of P_{aphA} and P_{tcpPH} allelic variants across the *V. cholerae* phylogeny. Isolates that were phenotypically serogroup O1 (for which those data were available) are indicated, as are those determined computationally to be serogroup O1. Lineages of *V. cholerae* O1 are indicated. Since *tcpPH* genes are harboured on VPI-1, P_{tcpPH} could only be detected in genomes containing VPI-1.

These data strongly indicate that genetic determinants which confer classical phenotypes (acetoin production and virulence gene expression) are almost exclusively in isolates that belong to the Classical lineage. Additional biotyping determinants were similarly studied, to determine whether this applies just to these promoters, or represents a more general observation.

5.3.8.3 – Polymyxin B sensitivity and haemolysis

El Tor biotype *V. cholerae* are classified as being resistant to polymyxin B – specifically, for an El Tor strain, a zone of inhibition will not be visible around a disc containing 50 units of polymyxin B [206]. In contrast, classical *V. cholerae* are sensitive to this concentration of polymyxin B and a zone of inhibition will be visible around the disc (the size of the zone of inhibition is not important in this specific assay) [206]. Briefly, the *almEFG* operon encodes three proteins which are responsible for catalysing the transfer of glycine residues to lipid A in El Tor *V. cholerae*, rendering them resistant to polymyxin B [461]. In a Classical isolate (O395), it was observed that the *almF* gene was disrupted, and that this mutation was responsible for the inability of this strain to resist polymyxin B – expression of an intact *almEFG* operon *in trans* dramatically elevated polymyxin B resistance in O395 [461].

Haemolysis in *V. cholerae* is mediated by the secreted haemolysin HlyA, which is processed into an active form by proteolysis [462, 463]. As mentioned in the Introduction (section 1.3.1.4), El Tor *V. cholerae* are characteristically haemolytic, and classical isolates are not [40, 206]. Classical isolates have been shown to be non-haemolytic as a consequence of a frameshift mutation in *hlyA* [464]. The hybrid strains of El Tor *V. cholerae* have been reported to be non-haemolytic, or haemolytic with a range of phenotypes [222]. Recent work suggests that multiple independent mutations in *hlyA* can be detected amongst hybrid *V. cholerae* [221]. It should also be noted that haemolysis is recognised to be an unreliable phenotype by which to characterise *V. cholerae* [206].

The state of *almEFG* and *hlyA* were determined across the phylogeny (Figure 5.21). Again, it was immediately evident that the mutations that confer a canonical classical biotyping phenotype – non-haemolysis due to an *hlyA* frameshift and sensitivity to polymyxin B due to an *almF* disruption – are both exclusive to the Classical lineage of *V. cholerae* (Figure 5.21).

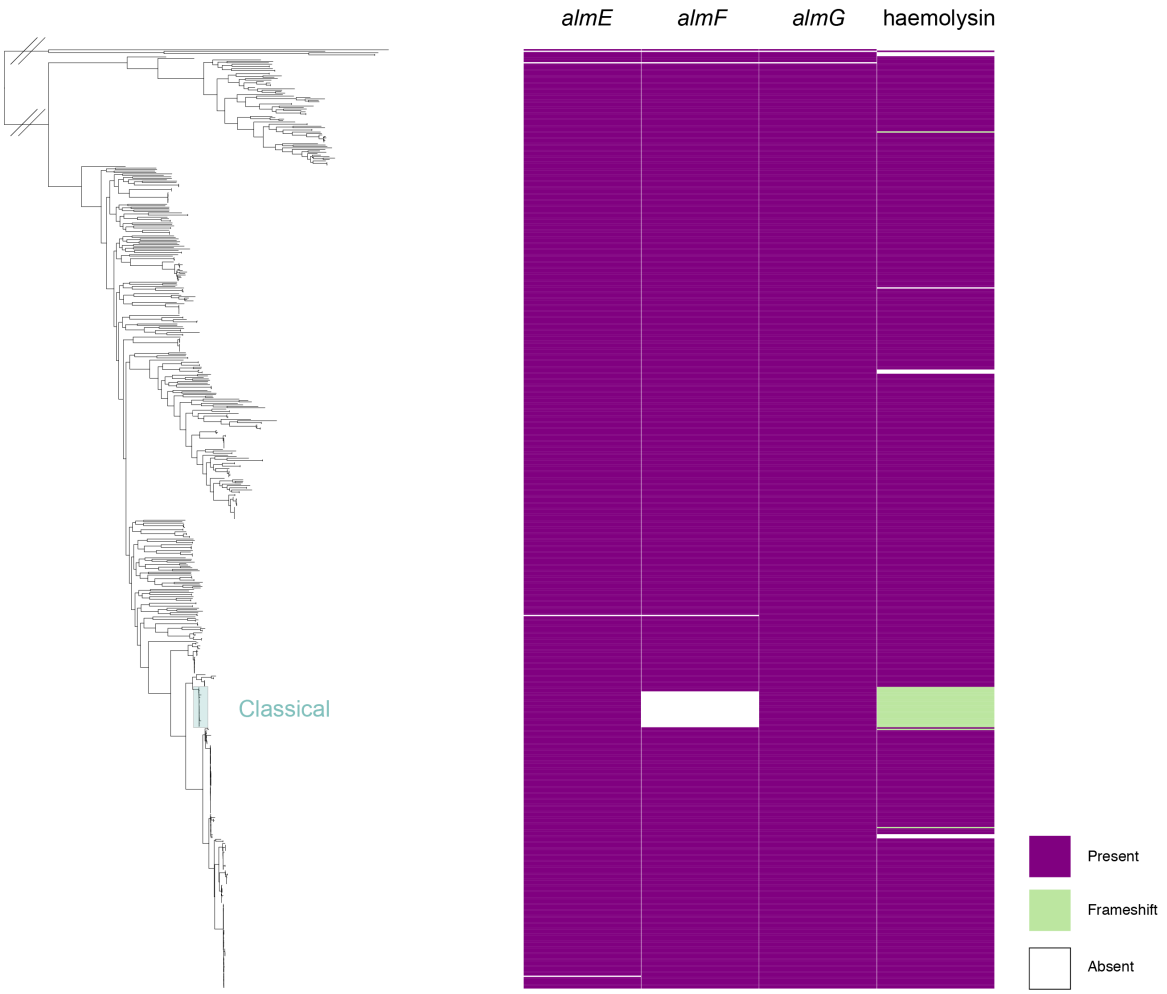


Figure 5.21 – Presence of intact and disrupted polymyxin B resistance genes and *hlyA* across the *V. cholerae* phylogeny. The absence of *almF* was determined from the pangenome matrix, as was the presence of the classical *hlyA* frameshift.

Collectively, these data show that genetic determinants which confer *V. cholerae* biotypes are, in fact, describing the Classical lineage rather than 7PET.

5.4 – Discussion

The results presented in this chapter describe some important aspects of *V. cholerae* biology, and challenge a number of dogmatic views in *V. cholerae* research. Firstly, it is not true to say that the VSP and VPI pathogenicity islands are found exclusively in epidemic or pandemic lineages of *V. cholerae*. Figures 5.5 and 5.7 show that these elements can be found in other *V. cholerae* of serogroup O1 as well as in non-O1 bacteria, consistent with previous reports (*e.g.*, [55, 396, 440]) and summarised below in Figure 5.22. Moreover, the presence of VPI-1 in non-O1 isolates underlines the fact that *V. cholerae* other than the members of pandemic lineages have the capacity to become lysogenised with CTXφ and to gain the capacity to express CT.

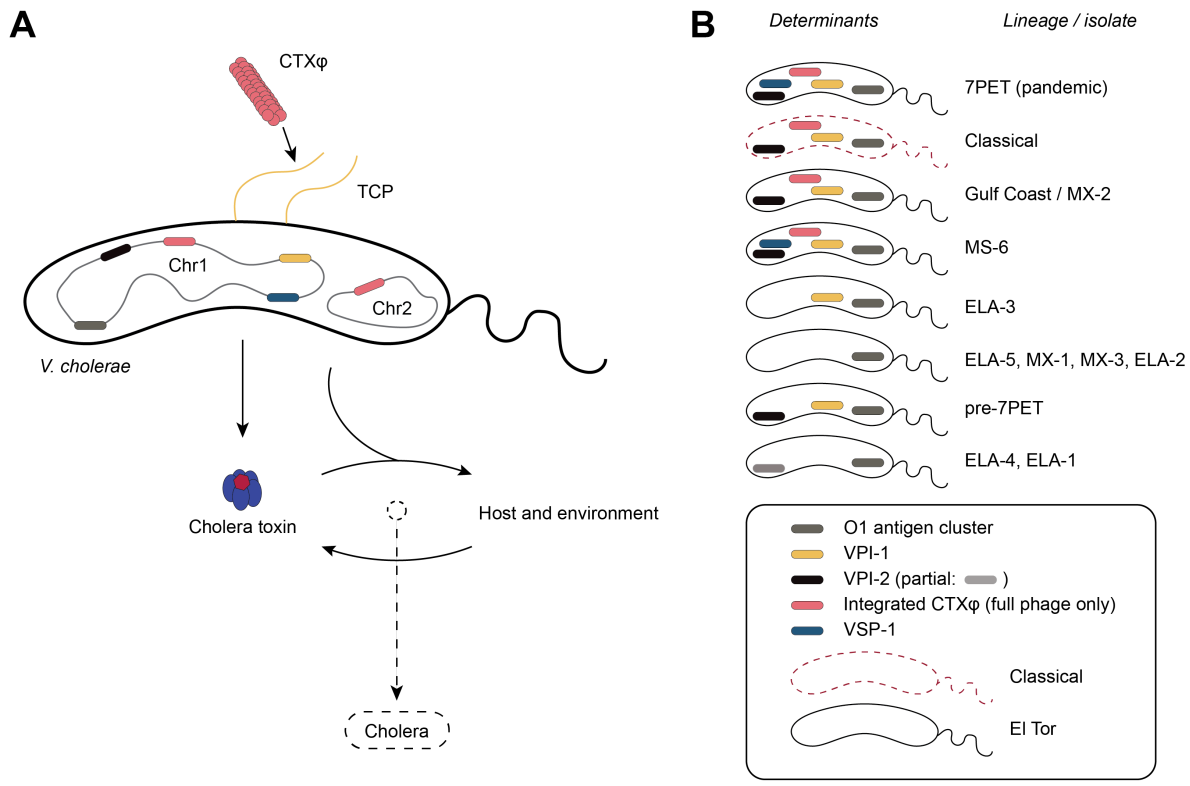


Figure 5.22 – Model of genetic determinants contributing to cholera. (A): Canonically, *V. cholerae* O1 harbouring a combination of genetic determinants can produce CT in a human host, causing cholera. However, the data presented in this chapter, as well as other work cited throughout, has shown that these genetic determinants can be found in multiple lineages of *V. cholerae* O1, not just those that cause pandemic cholera (B).

Our results suggest that the stocks of NCTC 8457 held by NCTC may not be the same as the strain that was sequenced in 2007 (Figure 5.10). The additional data available to us, including the fact that our sequenced stock is genomically of serogroup O1 and is phylogenetically

related to other non-7PET *V. cholerae* O1, and that this sequence appears to harbour a small extrachromosomal plasmid (Figures 5.13, 5.14), indicate that in order to resolve this satisfactorily, it will be necessary to obtain and re-sequence a culture of the originally-sequenced NCTC 8457 stock. Since this original sequence was generated by TIGR in the USA, it is possible that the stocks of NCTC 8457 held by ATCC (under accession # ATCC 14033; Table 5.2) and by NCTC are not identical.

To our knowledge, the data in this chapter represent the first report of a sequenced IncA/C2 plasmid which lacks any resistance determinants. Moreover, since NCTC 3661 was isolated in 1931, this suggests that pNCTC3661 is one of the oldest sequenced IncA/C2 plasmids and may be an example of an ancestral IncA/C2 plasmid backbone into which antimicrobial resistance determinants were imported as the use of antibiotics as therapeutics became more frequent. This sequence may be of considerable utility for the construction of plasmid phylogenies, or otherwise for understanding how IncA/C2 plasmids evolved to acquire AMR determinants over time. Since this plasmid type seems to be one of very few types that establish themselves in *V. cholerae* (Figures 5.1, 5.8; [158]), studying this plasmid type both within and outside the context of *Vibrio* spp. may be of considerable evolutionary interest.

This cluster of historical non-7PET serogroup O1 *V. cholerae* from NCTC underline the fact that using biotyping to identify *V. cholerae* of clinical importance is of very limited utility. These data show that all of the principal phenotypes upon which the biotyping scheme was devised describe the Classical lineage, rather than an El Tor lineage. This is exemplified by NCTC 9422 and the fact that this was co-deposited alongside other El Tor bacteria to which it was distantly related. It is apparent that the El Tor phenotype is much more broadly distributed across *V. cholerae* than just 7PET and local O1 lineages, whereas the specific mutations causing the Classical biotype appear to be coincident only in the Classical lineage. This genomic evidence is complementary to a conclusion from Chapter 3, in which the reliance on Inaba/Ogawa serotyping for epidemiological purposes was shown to be of very limited utility. Taken together, the work in this thesis so far strongly suggests that, in the light of genomic evidence, many of the bacteriological and biochemical assays used to study epidemic *V. cholerae* are misleading.

It has been observed that recent 7PET isolates, such as those which caused cholera outbreaks in Yemen, were sensitive to polymyxin B in spite of being part of an “El Tor” lineage [309]. It

was postulated that a non-synonymous mutation in *vprA* (causing a D89N mutation in VprA) was responsible for this phenotype, because functional VprA is required for the expression of the *almEFG* operon [309, 465]. This indicates that the acquisition of polymyxin B sensitivity in these recent 7PET isolates is not due to the acquisition of a Classical *almF* genotype – indeed, *almF* is intact in these recent 7PET isolates, a representative of which was included in the phylogeny discussed in this chapter (Figure 5.21). This indicates that although recent hybrid 7PET isolates appear to be acquiring "classical-like" phenotypes, this is not due to the acquisition of the same genetic mutations found in Classical isolates [221, 222, 309]. Understanding whether acquiring classical phenotypes alter the fitness of 7PET is an area of active research [466].

It appears that as diverse bacteria continue to be sequenced, the *V. cholerae* phylogeny continues to expand. Of particular interest for future study is the population of diverse genomes in the phylogeny, with a lower ANI to 7PET than the rest of the species (Figures 5.1, 5.3). It is possible that this clade corresponds to the 'novel species' or 'basal lineage' described by [467], though such a taxonomic determination would require considerable additional computational and microbiological study, and was outside the scope of this thesis research. As part of this PhD, the number of genomes available in this clade has been substantially expanded, and includes two closed genomes (NCTC 30, 48853_F01; Chapter 4). Thus, this is a promising area of future research into the diversity of the *V. cholerae* species.

In the final results chapter of this thesis, I consolidate the lines of enquiry that I have followed thus far, exploring the differences between pandemic and non-pandemic *V. cholerae*. To do this, I select eight key isolates from the phylogeny presented in this chapter, and design experiments by which to characterise the differences in the transcriptomes of these isolates when cultured under the same growth conditions. This is to explore whether gene expression patterns are more similar within or between lineages, and whether differences in regulation can be detected within lineages.

Chapter 6

Variation in gene expression in phylogenetically-selected *V. cholerae*

Contribution statement

Nick Thomson supervised the work described in this chapter. George Salmond, Gordon Dougan, and Julian Parkhill made substantive intellectual contributions to experimental design. rRNA depletion, library construction, and cDNA sequencing was performed by the WSI Bespoke RNA team. The live isolates used in this project were supplied by IVI, NCTC, and PHE, and have been described in Chapter 5.

I performed RNA-seq experiments, analysed the data and produced all figures.

COVID-19 statement

The work described in this chapter was severely affected by the shutdown imposed on the University of Cambridge and the Sanger Institute by the COVID-19 pandemic. This affected both the generation and analysis of RNA-seq data, as well as *in vitro* validation of the results from transcriptomic analysis. Aspects of the analysis herein may therefore be incomplete, but are the subject of current and future research.

6.1 – Overview

Throughout the preceding chapters, it has been demonstrated and discussed that the O1 serogroup and El Tor biotype do not exclusively describe the lineage of *V. cholerae* causing current pandemic cholera. It is also clear that pathogenicity islands and virulence determinants once thought to be found solely in pandemic *V. cholerae* are not exclusively found in pandemic lineages of *V. cholerae* (Chapter 5). In spite of this, the pandemic 7PET lineage continues to present a considerably elevated public health threat relative to that of local or endemic *V. cholerae*, as has been shown previously [189] and is exemplified by the Argentinian case study presented in Chapter 3. The biological basis of this phenomenon remains unclear.

Although *V. cholerae* can be compared and contrasted in terms of their gene complements (as has been done throughout this thesis), it is also necessary to consider functional differences between these bacteria. One approach that can be applied to understanding the biological differences between related bacteria is that of comparative transcriptomics, whereby bacteria are cultured under similar conditions and the expression profile of the genes they harbour is determined. Such experiments provide a snapshot of gene expression under appropriately-controlled *in vitro* (or *in vivo*) growth conditions.

Comparative transcriptomic experiments have been carried out in *V. cholerae* since the development of microarrays using the sequence of the N16961 reference genome [59]. Microarray-based experiments rely on the hybridisation of fluorescently-labelled cDNA, generated from total RNA extracted from *V. cholerae* grown under specific conditions, to spots of complementary ssDNA corresponding to genes on glass microarray slides. Experiments using microarrays have been used to characterise *V. cholerae* gene expression *in vitro* [468] and during infection [128], in the hyperinfectious state post-shedding [469], and to define regulons by comparing isogenic knock-outs in master regulators to a wild-type background [129]. This approach has also been used to characterise differences in gene expression between classical and El Tor biotype isolates (O395 and A1552, respectively) [470].

Over time, transcriptomic experiments shifted from microarray-based experiments to those employing RNA-seq, the use of next-generation sequencing to sequence quantitatively cDNA produced from *V. cholerae*. These experiments have included exploring bacterial gene expression during infection [471] and defining gene regulation networks [472, 473]. Recent

work has sought to collate previously-published *V. cholerae* gene expression studies for the purposes of meta-analyses [474]. By sequencing transcripts from the bacterium, RNA-seq also has the advantage of being able to monitor changes in expression of small RNAs (sRNAs) (*e.g.*, [475]), and other elements that microarray experiments cannot capture. It is important to note that the majority of transcriptomic experiments in *V. cholerae* have made use of a small number of laboratory strains. These include the 7PET strains N16961 and A1552, and the Classical strain O395.

In this chapter, I describe pilot experiments designed to optimise RNA-seq analysis and experiments for use under CL3 conditions. As mentioned in section 4.1, *V. cholerae* is classified as a bioterrorism agent under ATCSA Schedule 5. The Schedule 5-compliant laboratories at WSI operate at CL3, which imposes considerable constraint on the scale and scope of experimental work on *V. cholerae*. Therefore, preliminary experiments were important to optimise RNA isolation protocols under these containment conditions. As part of establishing these methodologies, I performed a transcriptomic experiment designed to explore the variation in gene expression of eight *V. cholerae* across multiple growth conditions, chosen to represent multiple lineages and genotypes of *V. cholerae*. I present the rationale for choosing these isolates, which was principally to determine whether pandemic and non-pandemic *V. cholerae* O1 exhibit different patterns of gene expression. I highlight a number of these results which are amenable to future study, and discuss how these follow-up experiments might be performed and what they might tell us about *V. cholerae* biology.

6.2 – Specific aims

In this final chapter, I aimed to:

1. Optimise experimental methods for CL3 operations by comparing the transcriptomes of Classical and 7PET *V. cholerae* grown under identical conditions, and compare the lists of differentially-expressed genes to those identified in previous studies,
2. Select rationally a total of eight live *V. cholerae* isolates suitable for transcriptomic comparisons, and
3. Execute a large pilot experiment to compare gene expression across eight *V. cholerae*, and to carry out an initial analysis of the resulting data.

6.3 – Results

6.3.1 – Methods optimisation and initial transcriptomic studies

Initial experiments were carried out in order to optimise transcriptomic methodologies for use in our CL3 laboratory. These experiments were also carried out in order to assess (i) the quality of RNA produced by our extraction protocol, (ii) the reproducibility within an experiment carried out under CL3 conditions, and (iii) the degree to which transcriptomic data generated using the live isolates at our disposal were concordant with previously-published datasets. In preliminary experiments, biological triplicate 25 ml cultures of *V. cholerae* were grown to an OD_{600} of 0.55 in LB liquid media at a defined temperature (either 30 or 37 ˚C) in baffled flasks with aeration (180 rpm). The OD_{600} of each culture was measured hourly. Once the target OD_{600} was reached, bacteria were collected from the culture by centrifugation, and the cell pellet frozen immediately at -80 ˚C. Total RNA was then extracted from cell pellets (details set out in Methods, section 2.2.16). This methodology is also outlined in Figure 6.1.

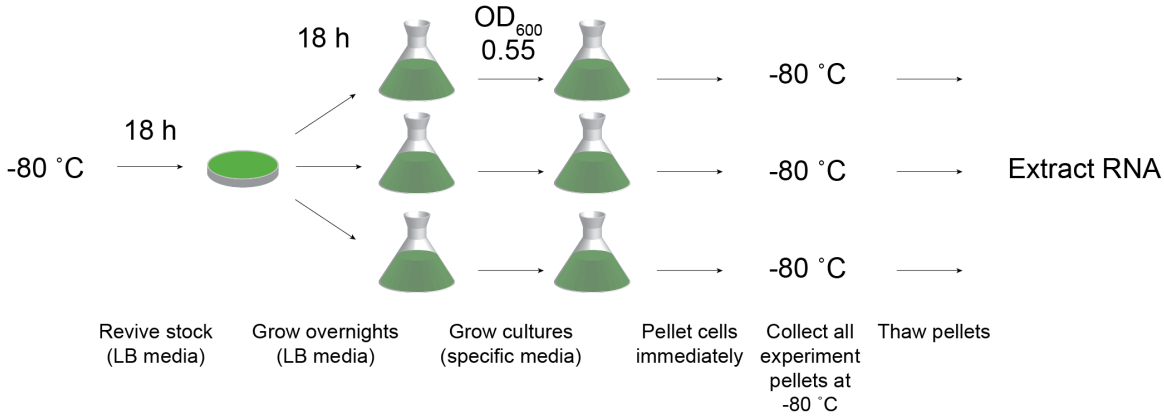


Figure 6.1 – Overview of RNA-seq experimental methodology for pilot experiment. All manipulations with the exception of culture incubation were performed inside a Class II MSC.

As well as assessing whether intact and sequence-able RNA could be isolated in the CL3 laboratory, this experiment aimed to determine whether differences in gene expression could be detected between a 7PET and a Classical culture grown to the same density. Cultures of bacteria were grown at both 37 ˚C and 30 ˚C, to assess the influence of temperature on gene expression in a strain, and whether these differences were similar or different across strains.

6.3.2 - RNA integrity and sequencing

Prior to submitting extracted RNA for rRNA depletion, library preparation and sequencing, the integrity of extracted RNA was assessed both by gel electrophoresis and by analysis on an Agilent Bioanalyser 2100 (Methods, section 2.2.17; data not shown). Only intact RNA samples were used for library construction; any which appeared to be degraded were discarded. Following the depletion of rRNA, strand-specific library preparation was performed and libraries were sequenced on one Illumina HiSeq 4000 lane. All of the samples for this experiment were batched together and sequenced on the same HiSeq lane, aiming to produce at least 10 million reads per sample. Details of total numbers of reads per sample, and the number of reads mapped to the N16961 genome sequence, are listed in Table 6.1 (see section 2.1.18 for computational details). Briefly, sequencing reads were mapped to the sequences of chromosomes 1 and 2 separately, gene expression was quantified, and differential gene expression analysis was performed using the Rockhopper package.

Sample ID	Strain	Condition	Number of reads	Reads mapped Chr1	Reads mapped Chr2
25900_7#1	7PET	37 ˚C, rep 1	14,092,482	12,200,919 (87%)	353,889 (3%)
25900_7#2	7PET	37 ˚C, rep 2	16,353,593	14,384,549 (88%)	283,329 (2%)
25900_7#3	7PET	37 ˚C, rep 3	22,897,140	19,977,810 (87%)	239,410 (1%)
25900_7#4	Classical	37 ˚C, rep 1	13,233,279	11,288,322 (85%)	334,318 (3%)
25900_7#5	Classical	37 ˚C, rep 2	18,441,714	15,733,617 (85%)	516,216 (3%)
25900_7#6	Classical	37 ˚C, rep 3	17,065,558	14,213,502 (83%)	730,105 (4%)
25900_7#19	7PET	30 ˚C, rep 1	16,724,553	12,862,968 (77%)	2,058,588 (12%)
25900_7#20	7PET	30 ˚C, rep 2	25,965,794	20,038,037 (77%)	3,092,668 (12%)
25900_7#21	7PET	30 ˚C, rep 3	26,594,182	21,059,504 (79%)	2,645,698 (10%)
25900_7#22	Classical	30 ˚C, rep 1	27,133,194	21,954,557 (81%)	1,552,086 (6%)
25900_7#23	Classical	30 ˚C, rep 2	20,970,321	16,216,919 (77%)	1,659,157 (8%)
25900_7#24	Classical	30 ˚C, rep 3	22,291,274	18,175,615 (82%)	1,026,833 (5%)

Table 6.1 – Lane IDs and summary statistics for RNA sequenced in pilot experiment. All reads mapped to the N16961 reference genome, accession # GCA_000006745.1. Strains chosen were MJD1402 (7PET) and MJD1404 (Classical); details presented in Table 6.6.

6.3.3 – Identification of differentially-expressed genes in pilot data

Previous studies of inter-biotype differential gene expression had been performed using microarray technologies, and had quantified differential gene expression for *V. cholerae* cultured in M9 + NRES minimal media at 30 ˚C to an OD_{600} of 0.65 [470]. Supplementation

of minimal media with asparagine, arginine, glutamate and serine (NRES) is required to induce expression of TCP and CT in minimal media [476]. This previous work had compared gene expression in the O395 strain of Classical lineage *V. cholerae* to that of A1552, a 7PET LAT-1 isolate from 1992 (detailed in section 3.4.5). Although the target OD_{600} and growth temperature in our pilot experiment was similar to those used by Beyhan *et al.*, the growth media were not identical.

Therefore, the first analysis carried out using these data aimed to identify genes which were differentially-expressed at 30 ˚C between 7PET and Classical *V. cholerae*. The fact that Beyhan and colleagues had performed a similar experiment using non-identical reference strains provided an important benchmark for our experiment, allowing for any expression differences that might be expected from the literature to be identified. Genes were determined to be differentially expressed if the $\log_2$ fold-change of expression between conditions was greater than or equal to 2, or less than or equal to -2, and that the q-value for that calculation was less than 0.01 (to account for multiple testing). Genes which met these criteria are listed in Tables 6.2 and 6.3 - triplicate data have been normalised to produce average expression values.

Gene ID	Product	Expression.7PET	Expression.Cl	pValue	qValue	log2Ratio
		Chromosome 1				
VC_0107	hypothetical protein	1323	209	9.07E-33	1.79E-31	2.6622 3821
VC_0163	conserved hypothetical protein	237	39	5.23E-15	7.21E-14	2.6033 4103
VC_0181	conserved hypothetical protein	339	1	0	0	8.4051 4146
VC_0205	hypothetical protein	141	0	0	0	Inf
VC_0285	4-hydroxy-2-oxoglutarate aldolase/2-deydro-3-deoxyphosphogluconate aldolase	146	8	2.09E-122	6.04E-121	4.1898 2456
VC_0471	SprT protein, putative	91	9	6.80E-27	1.26E-25	3.3378 6964
VC_0502	type IV pilin, putative	707	1	0	0	9.4655 664
VC_0512	methyl-accepting chemotaxis protein	331	0	0	0	Inf
VC_0707	hypothetical protein	1298	45	0	0	4.8502 2157
VC_0732	transcriptional regulator, LysR family	90	14	7.19E-19	1.12E-17	2.6844 9817
VC_0771	vibriobactin-specific isochorismatase	224	31	1.09E-43	2.45E-42	2.8531 5861
VC_0775	vibriobactin synthesis protein, putative	4864	460	1.18E-239	4.13E-238	3.4024 3746
VC_0867	hypothetical protein	132	23	4.96E-08	4.92E-07	2.5208 3216
VC_0990	transcriptional activator RfaH, putative	814	69	1.29E-154	4.01E-153	3.5603 6053
VC_1008	sodium-type flagellar protein MotY	980	94	9.33E-201	3.10E-199	3.3820 4909
VC_1292	hypothetical protein	4296	714	2.68E-53	6.41E-52	2.5889 9801
VC_1455	transcriptional repressor RstR	96	0	0	0	Inf
VC_1464	transcriptional repressor RstR	96	0	0	0	Inf
VC_1616	glutaredoxin, putative	290	15	2.93E-114	8.38E-113	4.2730 1849
VC_1654	hypothetical protein	160	13	2.48E-38	5.32E-37	3.6214 8838
VC_1788	hypothetical protein	135	16	2.57E-34	5.16E-33	3.0768 156
VC_1822	PTS system, fructose-specific IIABC component	268	32	7.56E-104	2.13E-102	3.0660 8919
VC_1824	PTS system, nitrogen regulatory IIA component, putative	900	140	5.47E-72	1.39E-70	2.6844 9817
VC_1828	conserved hypothetical protein	1246	77	0	0	4.0163 0181
VC_1829	hypothetical protein	455	62	7.41E-20	1.19E-18	2.8755 2642
VC_1970	benzoate transport protein	1354	92	0	0	3.8794 5007
VC_2147	hypothetical protein	173	28	1.23E-15	1.74E-14	2.6272 7331
VC_2324	transcriptional regulator, LysR family	96	4	5.67E-172	1.83E-170	4.5849 625
VC_2327	hypothetical protein	4128	205	0	0	4.3317 4716
VC_2328	hypothetical protein	260	25	7.67E-27	1.42E-25	3.3785 1162
VC_2329	2,3,4,5-tetrahydropyridine-2-carboxylate N-succinyltransferase	95	15	1.63E-21	2.75E-20	2.6629 6501
VC_2365	hypothetical protein	1352	190	1.52E-85	4.11E-84	2.8310 2383
VC_2509	hypothetical protein	93	8	2.01E-20	3.30E-19	3.5391 5881
VC_2534	magnesium transporter	216	27	4.18E-67	1.04E-65	3
VC_2544	fructose-1,6-bisphosphatase	106	7	4.93E-110	1.40E-108	3.9205 6553
VC_2545	inorganic pyrophosphatase	566	27	0	0	4.3897 7074
VC_2563	hypothetical protein	73	12	2.25E-12	2.83E-11	2.6048 6206
VC_2608	ABC transporter, ATP-binding protein	176	20	1.56E-85	4.19E-84	3.1375 0352

VC_2613	phosphoribulokinase	73	3	8.29E-207	2.78E-205	4.6048 6206
VC_2707	hypothetical protein	126	21	7.79E-09	8.09E-08	2.5849 625
VC_2759	fatty oxidation complex, beta subunit	143	20	1.30E-39	2.81E-38	2.8379 4324
VC_t038	tRNA-Met	186	15	5.27E-35	1.07E-33	3.6322 6822
VC_t068	tRNA-Leu	95	14	4.27E-09	4.57E-08	2.7625 0069
VC_0107	hypothetical protein	1323	209	9.07E-33	1.79E-31	2.6622 3821
VC_0163	conserved hypothetical protein	237	39	5.23E-15	7.21E-14	2.6033 4103
VC_0181	conserved hypothetical protein	339	1	0	0	8.4051 4146
VC_0205	hypothetical protein	141	0	0	0	Inf
VC_0285	4-hydroxy-2-oxoglutarate aldolase/2-deydro-3-deoxyphosphogluconate aldolase	146	8	2.09E-122	6.04E-121	4.1898 2456
VC_0471	SprT protein, putative	91	9	6.80E-27	1.26E-25	3.3378 6964
VC_0502	type IV pilin, putative	707	1	0	0	9.4655 664
VC_0512	methyl-accepting chemotaxis protein	331	0	0	0	Inf
VC_0707	hypothetical protein	1298	45	0	0	4.8502 2157
VC_0732	transcriptional regulator, LysR family	90	14	7.19E-19	1.12E-17	2.6844 9817
VC_0771	vibriobactin-specific isochorismatase	224	31	1.09E-43	2.45E-42	2.8531 5861
VC_0775	vibriobactin synthesis protein, putative	4864	460	1.18E-239	4.13E-238	3.4024 3746
VC_0867	hypothetical protein	132	23	4.96E-08	4.92E-07	2.5208 3216
VC_0990	transcriptional activator RfaH, putative	814	69	1.29E-154	4.01E-153	3.5603 6053
VC_1008	sodium-type flagellar protein MotY	980	94	9.33E-201	3.10E-199	3.3820 4909
VC_1292	hypothetical protein	4296	714	2.68E-53	6.41E-52	2.5889 9801
VC_1455	transcriptional repressor RstR	96	0	0	0	Inf
VC_1464	transcriptional repressor RstR	96	0	0	0	Inf
VC_1616	glutaredoxin, putative	290	15	2.93E-114	8.38E-113	4.2730 1849
VC_1654	hypothetical protein	160	13	2.48E-38	5.32E-37	3.6214 8838
VC_1788	hypothetical protein	135	16	2.57E-34	5.16E-33	3.0768 156
VC_1822	PTS system, fructose-specific IIABC component	268	32	7.56E-104	2.13E-102	3.0660 8919
VC_1824	PTS system, nitrogen regulatory IIA component, putative	900	140	5.47E-72	1.39E-70	2.6844 9817
VC_1828	conserved hypothetical protein	1246	77	0	0	4.0163 0181
	Chromosome 2					
VC_A0029	transcriptional regulator, putative	96	21	8.07E-08	4.60E-07	2.1926 4508
VC_A0077	sulfate permease family protein	188	40	2.35E-22	2.27E-21	2.2326 6076
VC_A0080	GGDEF family protein	360	38	8.89E-121	1.72E-119	3.2439 2558
VC_A0106	hypothetical protein	92	20	1.44E-10	9.50E-10	2.2016 3386
VC_A0224	hypothetical protein	92	11	2.21E-16	1.78E-15	3.0641 3034
VC_A0233	hypothetical protein	199	37	4.92E-10	3.10E-09	2.4271 7125
VC_A0296	hypothetical protein	4114	502	2.33E-70	3.91E-69	3.0347 8252
VC_A0298	hypothetical protein	76	8	3.62E-16	2.89E-15	3.2479 2751
VC_A0310	hypothetical protein	174	2	0	0	6.4429 435
VC_A0321	hypothetical protein	544	27	5.53E-113	1.00E-111	4.3325 7534

VC_A0329	hypothetical protein	93	8	7.91E-25	7.88E-24	3.5391 5881
VC_A0335	hypothetical protein	4685	516	2.46E-91	4.32E-90	3.1826 0608
VC_A0348	conserved hypothetical protein	694	31	1.05E-272	2.79E-271	4.4845 9554
VC_A0352	hypothetical protein	2113	316	3.36E-31	3.87E-30	2.7412 963
VC_A0359	plasmid stabilization element ParE, putative	1014	245	3.39E-18	2.89E-17	2.0492 04
VC_A0362	hypothetical protein	7734	1691	2.26E-69	3.76E-68	2.1933 381
VC_A0369	hypothetical protein	361	28	1.33E-76	2.27E-75	3.6885 001
VC_A0370	hypothetical protein	224	22	7.34E-52	1.06E-50	3.3479 233
VC_A0424	hypothetical protein	250	44	1.02E-15	8.12E-15	2.5063 5267
VC_A0429	hypothetical protein	2316	562	2.15E-15	1.68E-14	2.0429 9322
VC_A0432	hypothetical protein	75	10	2.61E-14	2.02E-13	2.9068 906
VC_A0434	hypothetical protein	3343	756	1.28E-25	1.32E-24	2.1446 8522
VC_A0448	hypothetical protein	104	14	1.04E-16	8.55E-16	2.8930 848
VC_A0466	hypothetical protein	2733	11	0	0	7.9568 3813
VC_A0503	conserved hypothetical protein	682	130	1.95E-18	1.69E-17	2.3912 6012
VC_A0515	hypothetical protein	3054	292	3.24E-92	5.75E-91	3.3866 5979
VC_A0531	sensor histidine kinase	336	47	1.36E-62	2.12E-61	2.8377 2857
VC_A0620	thiosulfate sulfurtransferase SseA, putative	125	30	3.14E-09	1.93E-08	2.0588 9369
VC_A0685	iron(III) ABC transporter, periplasmic iron-compound-binding protein	203	27	6.24E-37	7.90E-36	2.9104 4842
VC_A0686	iron(III) ABC transporter, permease protein	95	14	9.12E-29	9.99E-28	2.7625 0069
VC_A0687	iron(III) ABC transporter, ATP-binding protein	113	4	2.36E-239	5.97E-238	4.8201 7896
VC_A0713	hypothetical protein	73	6	2.46E-19	2.23E-18	3.6048 6206
VC_A0732	conserved hypothetical protein	152	3	0	0	5.6629 6501
VC_A0777	conserved hypothetical protein	263	50	3.00E-18	2.58E-17	2.3950 628
VC_A0874	hypothetical protein	1126	103	8.28E-79	1.44E-77	3.4504 9058
VC_A0879	hypothetical protein	2435	30	0	0	6.3428 1546
VC_A0892	hypothetical protein	101	24	1.25E-05	5.95E-05	2.0732 4898
VC_A0928	hypothetical protein	219	36	1.33E-17	1.12E-16	2.6048 6206
VC_A0945	maltose ABC transporter, periplasmic maltose-binding protein	119	24	3.76E-14	2.89E-13	2.3098 5526
VC_A0978	amino acid ABC transporter, periplasmic amino acid-binding protein, putative	100	21	1.25E-08	7.49E-08	2.2515 3877
VC_A0997	hypothetical protein	65	11	1.84E-12	1.32E-11	2.5629 3619
VC_A1007	hypothetical protein	66	5	2.86E-31	3.32E-30	3.7224 6602
VC_A1028	maltoporin	114	28	3.05E-10	1.95E-09	2.0255 3509

Table 6.2 - Genes upregulated in 7PET relative to Classical (*i.e.*, $\log_2$FC ≥ 2) at 30 ˚C. Genes that were previously identified as differentially-regulated between 7PET and Classical isolates at 30 ˚C [470] are indicated in blue. Two copies of *rstR* are present in the N16961 reference genome (Figure 4.2).

Gene ID	Product	Expression.7PET	Expression.Cl	pValue	qValue	log2Ratio
	Chromosome 1					
VC_0160	hypothetical protein	91	659	2.70E-37	5.68E-36	-2.85634
VC_0198	hypothetical protein	44	712	0	0	-4.0163018
VC_0204	conserved hypothetical protein	30	236	3.28E-85	8.65E-84	-2.9757525
VC_0292	hypothetical protein	40	168	7.63E-07	6.72E-06	-2.0703893
VC_0294	hypothetical protein	21	147	1.05E-16	1.54E-15	-2.8073549
VC_0314	hypothetical protein	843	7664	1.49E-150	4.59E-149	-3.184493
VC_0388	hypothetical protein	50	964	1.23E-251	4.34E-250	-4.2690331
VC_0458	conserved hypothetical protein	31	171	2.14E-17	3.21E-16	-2.4636562
VC_0485	pyruvate kinase I	101	464	1.10E-25	2.01E-24	-2.1997695
VC_0621	hypothetical protein	85	416	3.70E-31	7.10E-30	-2.2910488
VC_0713	hypothetical protein	768	6871	1.12E-135	3.34E-134	-3.1613419
VC_0765	conserved hypothetical protein	69	351	2.38E-22	4.12E-21	-2.3468028
VC_0824	TagD protein	99	765	5.50E-127	1.60E-125	-2.9499593
VC_0829	toxin co-regulated pilus biosynthesis protein B	16	78	3.73E-24	6.62E-23	-2.2854022
VC_0830	toxin co-regulated pilus biosynthesis protein Q	2	132	0	0	-6.0443941
VC_0831	toxin co-regulated pilus biosynthesis outer membrane protein C	20	109	4.16E-37	8.69E-36	-2.4462562
VC_0832	toxin co-regulated pilus biosynthesis protein R	9	195	3.98E-282	1.47E-280	-4.4374053
VC_0833	toxin co-regulated pilus biosynthesis protein D	4	182	0	0	-5.5077946
VC_0834	toxin co-regulated pilus biosynthesis protein S	9	340	0	0	-5.2394659
VC_0835	toxin co-regulated pilus biosynthesis protein T	13	156	2.64E-177	8.59E-176	-3.5849625
VC_0837	toxin co-regulated pilus biosynthesis protein F	43	1484	0	0	-5.1090106
VC_0838	TCP pilus virulence regulatory protein	5	66	5.95E-80	1.53E-78	-3.722466
VC_0839	leader peptidase TcpJ	16	83	7.15E-18	1.08E-16	-2.3750394
VC_0841	accessory colonization factor AcfC	26	330	2.31E-225	7.89E-224	-3.6658825
VC_0843	TagE protein	39	2266	0	0	-5.8605299
VC_0844	accessory colonization factor AcfA	14	968	0	0	-6.1115083
VC_0939	hypothetical protein	16	97	1.19E-18	1.83E-17	-2.5999128
VC_1233	conserved hypothetical protein	15	68	7.53E-09	7.86E-08	-2.1805722
VC_1290	DNA polymerase III, epsilon subunit, putative	49	211	5.01E-24	8.86E-23	-2.1063893
VC_1324	hypothetical protein	595	60038	0	0	-6.6568424
VC_1325	galactoside ABC transporter, periplasmic D-galactose/D-glucose-binding protein	12	365	0	0	-4.9267902
VC_1326	hypothetical protein	491	37233	0	0	-6.244715
VC_1327	galactoside ABC transporter, ATP-binding protein	8	80	3.10E-85	8.23E-84	-3.3219281
VC_1456	cholera enterotoxin, B subunit	34	465	4.56E-184	1.50E-182	-3.7736241
VC_1457	cholera enterotoxin, A subunit	27	462	0	0	-4.0968615
VC_1459	accessory cholera enterotoxin	78	455	5.86E-40	1.27E-38	-2.5443205
VC_1462	RstB2 protein	70	435	1.68E-49	3.94E-48	-2.6355886
VC_1463	RstA2 protein	24	101	2.61E-19	4.12E-18	-2.073249
VC_1554	glycerophosphoryl diester phosphodiesterase, putative	73	296	1.08E-28	2.04E-27	-2.0196288
VC_1569	hypothetical protein	18	86	4.68E-09	4.98E-08	-2.2563398
VC_1593	GGDEF family protein	91	665	8.83E-71	2.22E-69	-2.8694159
VC_1650	collagenase	56	320	4.66E-41	1.02E-39	-2.5145732
VC_1772	hypothetical protein	1016	13882	0	0	-3.7722431
VC_1774	conserved hypothetical protein	14	82	9.52E-32	1.84E-30	-2.5501971
VC_1776	N-acetylneuraminate lyase, putative	23	111	2.25E-23	3.93E-22	-2.2708539
VC_1780	hypothetical protein	271	7815	0	0	-4.8498811
VC_1825	transcriptional regulator	31	135	4.04E-21	6.77E-20	-2.1226193
VC_2083	zinc ABC transporter, permease protein	88	482	3.96E-65	9.82E-64	-2.4534577
VC_2306	hypothetical protein	1285	7697	7.70E-44	1.74E-42	-2.5825279
VC_2495	hypothetical protein	59	634	3.80E-73	9.70E-72	-3.425696

VC_2752	hypothetical protein	54	794	1.24E-143	3.78E-142	-3.8781077
		Chromosome 2				
VC_A0086	hypothetical protein	53	559	2.81E-117	5.29E-116	-3.398784
VC_A0087	hypothetical protein	118	824	3.51E-43	4.67E-42	-2.8038575
VC_A0128	ribose ABC transporter, ATP-binding protein	6	39	7.45E-21	6.97E-20	-2.7004397
VC_A0130	ribose ABC transporter, periplasmic D-ribose-binding protein	21	147	8.92E-42	1.16E-40	-2.8073549
VC_A0163	conserved hypothetical protein	26	142	1.99E-12	1.41E-11	-2.4493074
VC_A0480	hypothetical protein	287	1287	2.26E-29	2.49E-28	-2.1648894
VC_A0544	conserved hypothetical protein	114	665	1.09E-65	1.75E-64	-2.5443205
VC_A0582	conserved hypothetical protein	36	156	2.10E-13	1.57E-12	-2.1154772
VC_A0587	conserved hypothetical protein	36	206	4.04E-20	3.76E-19	-2.5165755
VC_A0665	C4-dicarboxylate transporter, anaerobic	108	536	3.06E-43	4.11E-42	-2.3112017
VC_A0670	hypothetical protein	230	186143	0	0	-9.6605618
VC_A0671	hypothetical protein	38	223	4.11E-51	5.87E-50	-2.5529724
VC_A0707	regulatory protein UhpC, putative	15	60	3.02E-12	2.13E-11	-2
VC_A0743	conserved hypothetical protein	7	451	0	0	-6.0096287
VC_A0744	glycerol kinase	45	186	4.20E-28	4.54E-27	-2.0473057
VC_A0750	hypothetical protein	237	5266	0	0	-4.4737486
VC_A0857	hypothetical protein	13	137	1.51E-30	1.71E-29	-3.3975924
VC_A0914	hemin ABC transporter, permease protein, putative	30	157	1.96E-30	2.21E-29	-2.3877302
VC_A0915	hemin ABC transporter, ATP-binding protein HutD	340	1652	9.95E-36	1.22E-34	-2.280607
VC_A0986	conserved hypothetical protein	88	474	1.99E-53	2.93E-52	-2.4293116
VC_A0989	conserved hypothetical protein	139	1238	2.78E-185	6.31E-184	-3.1548545
VC_A1003	hypothetical protein	344	1856	5.02E-51	7.12E-50	-2.4317162
VC_A1041	phosphomannomutase, putative	373	1896	1.10E-27	1.17E-26	-2.3457114
VC_A1062	putrescine-ornithine antiporter	9	36	7.85E-08	4.49E-07	-2

Table 6.3 - Genes upregulated in Classical relative to 7PET (*i.e.*, $\log_2$FC ≥ 2) at 30 ˚C. Genes that were previously identified as differentially-regulated between 7PET and Classical isolates at 30 ˚C [470] are indicated in blue. rRNA loci have been deleted from these results.

In each sample, the vast majority of reads mapped to loci on chromosome 1 rather than chromosome 2 (Table 6.1), suggesting that chromosome 1 is more transcriptionally active than chromosome 2. A total of 110 genes were found to be significantly elevated in expression in 7PET relative to Classical (67 on chromosome 1 and 43 on chromosome 2) and 82 genes were significantly elevated in expression in Classical relative to 7PET (58 on chromosome 1, 24 on chromosome 2) (Tables 6.2, 6.3). However, the overlap between these results and those of Beyhan and colleagues was limited. Behyan *et al* identified 270 genes that were elevated in expression in their Classical strain relative to their El Tor strain, and 252 genes that were elevated in their El Tor strain relative to their Classical strain. However, of these genes, just 11 overlap with our results from 7PET, and 16 overlap with our Classical isolate (Tables 6.2, 6.3). This may reflect strain-specific differences, the differences in growth media between the experiments, or differences in sensitivity between microarray-based analyses and this RNA-seq experiment.

In spite of this, it is notable that the differential expression of genes encoded by CTXφ and VPI-1 between 7PET and Classical was recapitulated in our experiment. This included the expression of the *rstR* gene in 7PET (encoding the RstR repressor of CTXφ genes, and under "normal growth conditions", this is the only CTXφ protein produced by lysogens [99]) and the expression of *ctxAB*, *tcp*, and *acf* virulence genes in Classical cultures under these *in vitro* conditions. This is consistent with previous studies discussed in Chapter 5 (section 5.3.8.2) which showed that genetic determinants such as P_{aphA} and P_{tcpPH} mutations drive overexpression of the ToxT regulon in Classical bacteria *in vitro*. Further experimentation and validation is required to determine whether the genes that were differentially regulated in this study but not in that of Beyhan *et al* are due to differences in culture conditions between these studies, or whether they reflect strain-specific variation in gene expression.

We then determined the sets of genes that were differentially expressed by these strains when cultured at 37 ˚C, rather than at 30 ˚C. All other experimental conditions were kept constant with the 30 ˚C experiment. The results are illustrated in volcano plots (Figure 6.2) and significantly differentially regulated genes are listed in Tables 6.4 and 6.5.

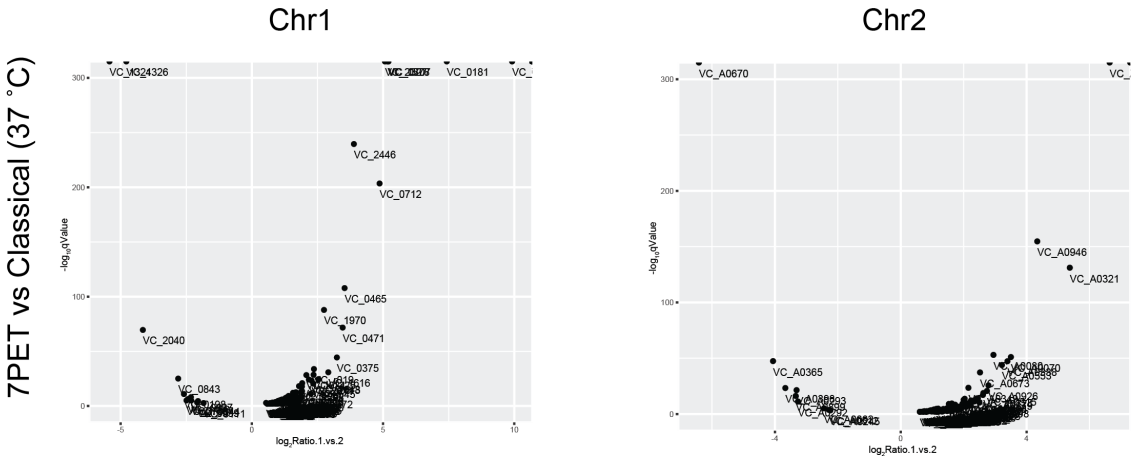


Figure 6.2 – Volcano plots comparing gene expression in 7PET and Classical strains at 37 ˚C. Genes that were not detected in one strain or the other are not represented in these plots (log_2FC = infinite). Genes which passed statistical thresholds described in section 2.1.18 are listed in Tables 6.4 and 6.5.

Gene ID	Product	Expression.7 PET	Expressio n.Cl	pValue	qValue	log2Rati o
		Chromosome 1				
VC_0175	deoxycytidylate deaminase-related protein	185	0	0	0	Inf
VC_0176	transcriptional regulator, putative	105	0	0	0	Inf
VC_0179	hypothetical protein	72	0	0	0	Inf
VC_0180	conserved hypothetical protein	87	0	0	0	Inf
VC_0181	conserved hypothetical protein	172	1	0	0	7.42626 4755
VC_0182	hypothetical protein	212	0	0	0	Inf
VC_0183	hypothetical protein	131	0	0	0	Inf
VC_0184	hypothetical protein	223	0	0	0	Inf
VC_0185	transposase, putative	204	0	0	0	Inf
VC_0375	hypothetical protein	236	25	1.22E-46	4.07E-45	3.23878 686
VC_0465	tyrosyl-tRNA synthetase	81	7	3.61E-110	1.50E-108	3.53249 5081
VC_0471	SprT protein, putative	88	8	5.14E-74	1.90E-72	3.45943 1619
VC_0490	conserved hypothetical protein	133	0	0	0	Inf
VC_0491	hypothetical protein	99	0	0	0	Inf
VC_0492	hypothetical protein	120	0	0	0	Inf
VC_0493	hypothetical protein	130	0	0	0	Inf
VC_0495	conserved hypothetical protein	164	0	0	0	Inf
VC_0496	hypothetical protein	77	0	0	0	Inf
VC_0502	type IV pilin, putative	345	0	0	0	Inf
VC_0503	conserved hypothetical protein	175	0	0	0	Inf
VC_0504	hypothetical protein	1705	0	0	0	Inf
VC_0505	hypothetical protein	966	1	0	0	9.91587 9379
VC_0506	hypothetical protein	529	0	0	0	Inf
VC_0507	hypothetical protein	359	0	0	0	Inf
VC_0508	hypothetical protein	143	0	0	0	Inf
VC_0509	hypothetical protein	112	0	0	0	Inf
VC_0510	DNA repair protein RadC-related protein	104	0	0	0	Inf
VC_0512	methyl-accepting chemotaxis protein	119	0	0	0	Inf
VC_0516	phage integrase	110	0	0	0	Inf
VC_0707	hypothetical protein	369	10	0	0	5.20554 8911
VC_0712	hypothetical protein	175	6	7.54E-206	3.43E-204	4.86624 8611
VC_1225	hypothetical protein	719	141	8.88E-31	2.61E-29	2.35029 6608
VC_1452	RstC protein	256	0	0	0	Inf
VC_1572	hypothetical protein	76	16	2.03E-13	4.89E-12	2.24792 7513
VC_1578	hypothetical protein	108	3	0	0	5.16992 5001
VC_1616	glutaredoxin, putative	143	19	4.80E-33	1.42E-31	2.91194 3823
VC_1648	hypothetical protein	633	109	8.75E-27	2.50E-25	2.53787 7365
VC_1970	benzoate transport protein	524	78	3.61E-90	1.41E-88	2.74802 0783
VC_2327	hypothetical protein	1983	59	0	0	5.07082 5913
VC_2446	tRNA nucleotidyltransferase	222	15	6.09E-242	2.93E-240	3.88752 5271
VC_2545	inorganic pyrophosphatase	172	34	1.31E-22	3.55E-21	2.33880 1913
VC_2566	conserved hypothetical protein	1526	339	2.36E-26	6.63E-25	2.17039 7784
VC_2611	hypothetical protein	1410	291	1.92E-25	5.32E-24	2.27660 4104
VC_2613	phosphoribulokinase	87	15	2.27E-26	6.43E-25	2.53605 29
		Chromosome 2				
VC_A0028	hypothetical protein	179	39	2.31E-11	2.93E-10	2.19841 356

VC_A0070	phosphate ABC transporter, periplasmic phosphate-binding protein	68	6	2.91E-53	8.46E-52	3.50250034
VC_A0080	GGDEF family protein	162	21	3.38E-55	1.00E-53	2.94753258
VC_A0168	hypothetical protein	164	36	7.49E-12	9.64E-11	2.187627
VC_A0315	hypothetical protein	799	8	0	0	6.64205169
VC_A0321	hypothetical protein	83	2	1.87E-133	7.60E-132	5.37503943
VC_A0323	conserved hypothetical protein	148	0	0	0	Inf
VC_A0328	biphenyl-2,3-diol 1,2-dioxygenase III-related protein	75	16	1.84E-09	2.04E-08	2.22881869
VC_A0335	hypothetical protein	957	233	1.12E-10	1.35E-09	2.03818897
VC_A0338	conserved hypothetical protein	63	6	1.75E-49	4.74E-48	3.39231742
VC_A0348	conserved hypothetical protein	263	49	5.93E-16	9.35E-15	2.42420915
VC_A0349	RelB protein	2128	479	1.46E-25	3.09E-24	2.15140059
VC_A0515	hypothetical protein	448	67	1.02E-22	2.07E-21	2.74126573
VC_A0527	conserved hypothetical protein	95	19	7.37E-14	1.06E-12	2.32192809
VC_A0531	sensor histidine kinase	131	32	1.42E-15	2.17E-14	2.033423
VC_A0553	hypothetical protein	223	24	4.55E-46	1.19E-44	3.2159374
VC_A0660	hypothetical protein	85	20	2.55E-08	2.66E-07	2.08746284
VC_A0673	transcriptional regulator, LacI family	373	65	2.15E-39	5.35E-38	2.52066401
VC_A0686	iron(III) ABC transporter, permease protein	84	21	1.24E-14	1.86E-13	2
VC_A0698	hypothetical protein	102	18	4.29E-12	5.62E-11	2.50250034
VC_A0719	sensor histide kinase	124	30	5.04E-12	6.54E-11	2.04730571
VC_A0774	UDP-glucose 4-epimerase	105	17	7.89E-21	1.54E-19	2.62678268
VC_A0879	hypothetical protein	159	26	4.26E-19	7.83E-18	2.61244324
VC_A0924	conserved hypothetical protein	66	14	4.92E-14	7.16E-13	2.2370392
VC_A0926	transcriptional regulator, AraC/XylS family	76	11	9.84E-28	2.17E-26	2.78849589
VC_A0946	maltose/maltodextrin ABC transporter, ATP-binding protein	81	4	4.12E-157	1.87E-155	4.33985
VC_A0994	hypothetical protein	93	17	6.80E-16	1.06E-14	2.45169597
VC_A1061	hypothetical protein	127	25	3.09E-13	4.42E-12	2.3448285

Table 6.4 - Genes upregulated in 7PET relative to Classical (*i.e.*, $\log_2FC \geq 2$) at 37 ˚C. Note that pathogenicity islands VSP-1 (*VC_0174-0186*) and VSP-2 (*VC_0489-0517*) are present and transcribed in 7PET isolates, but not in Classical isolates. rRNA and tRNA loci have been deleted from these results.

Gene ID	Product	Expression .7PET	Expression .Cl	pValue	qValue	log2Ratio
		Chromosome 1				
VC_0198	hypothetical protein	76	457	4.80E-13	1.15E-11	-2.588122842
VC_0733	hypothetical protein	16	90	4.35E-07	7.64E-06	-2.491853096
VC_0837	toxin co-regulated pilus biosynthesis protein F	47	235	3.41E-10	7.35E-09	-2.321928095

VC_0843	TagE protein	84	588	2.90E-27	8.39E-26	-2.807354922
VC_0844	accessory colonization factor AcfA	61	253	1.80E-06	2.91E-05	-2.052256237
VC_0939	hypothetical protein	37	157	4.23E-05	0.00057203	-2.085167383
VC_1324	hypothetical protein	118	5072	0	0	-5.425695981
VC_1326	hypothetical protein	161	4439	0	0	-4.785102115
VC_1456	cholera enterotoxin, B subunit	17	92	2.44E-08	4.69E-07	-2.436099115
VC_1457	cholera enterotoxin, A subunit	23	114	4.96E-07	8.62E-06	-2.309328058
VC_2040	conserved hypothetical protein	5	89	7.88E-72	2.84E-70	-4.153805336
	Chromosome 2					
VC_A0245	PTS system, IIA component	14	66	7.39E-05	0.00051408	-2.2370392
VC_A0292	hypothetical protein	8	77	4.50E-13	6.26E-12	-3.2667865
VC_A0293	hypothetical protein	13	129	1.85E-23	3.81E-22	-3.3107875
VC_A0365	hypothetical protein	6	100	1.36E-49	3.74E-48	-4.0588937
VC_A0398	hypothetical protein	7	89	2.35E-25	4.91E-24	-3.6683785
VC_A0512	anaerobic ribonucleoside-triphosphate reductase activating protein	21	102	2.68E-05	0.00020681	-2.2801079
VC_A0662	CbbY family protein	14	76	7.44E-07	7.14E-06	-2.4405726
VC_A0670	hypothetical protein	281	23993	0	0	-6.4158996
VC_A0899	hypothetical protein	18	182	3.46E-18	6.15E-17	-3.3378696

Table 6.5 - Genes upregulated in Classical relative to 7PET (*i.e.*, $\log_2$FC ≤ 2) at 37 ˚C. Genes encoded on VPI-1 and CTXφ were up-regulated (*VC_0809-VC_0848* and *VC_1457-1465*, respectively).

Once again, virulence genes that are members of the ToxT regulon were expressed in Classical at a higher level than in 7PET at 37 ˚C, consistent with the observations made at 30 ˚C. This included *ctxAB*, *tcp* and *acf* genes, and the *tagE* gene. It was also evident that at 37 ˚C, genes encoded by VSP-1 and VSP-2 were transcribed in 7PET (Table 6.4). These genes are absent from Classical *V. cholerae* and from MJD1402 (Figures 5.7, 5.22), and therefore their expression was detected as having an infinite $\log_2$(expression ratio). These genes were not determined to be significantly transcribed at 30 ˚C (Tables 6.2, 6.3). We also noted that in spite of rRNA depletion using RiboZero (Methods, section 2.2.18), rRNA loci were identified as being transcribed at significantly different levels in 7PET and Classical isolates (data not shown). This indicates that RiboZero depletion in these samples, or across the experiment more generally, was not totally efficient, and emphasises that this is a depletion rather than an elimination step.

6.3.4 – Effect of temperature on transcriptome of Classical and 7PET isolates

Using these data, a comparison was then performed to identify genes that were differentially expressed at different temperatures (*i.e.*, comparing the 37 ˚C and 30 ˚C results for one strain to one another). At 37 ˚C, 122 genes were upregulated in 7PET relative to Classical (81 on chromosome 1, 41 on chromosome 2) and 245 genes were upregulated in Classical relative to 7PET (154 on chromosome 1, 91 on chromosome 2). At 30 ˚C, 93 genes were upregulated in 7PET relative to Classical (59 on chromosome 1, 34 on chromosome 2), and 105 genes were upregulated in Classical relative to 7PET (67 on chromosome 1, 38 on chromosome 2). Sixty-two genes were upregulated in 7PET relative to Classical at both 37 and 30 ˚C (42 on chromosome 1, 20 on chromosome 2), and 32 genes were upregulated in Classical relative to 7PET at both temperatures (21 on chromosome 1, 11 on chromosome 2). These results are summarised in Figure 6.3; the raw data are not shown.

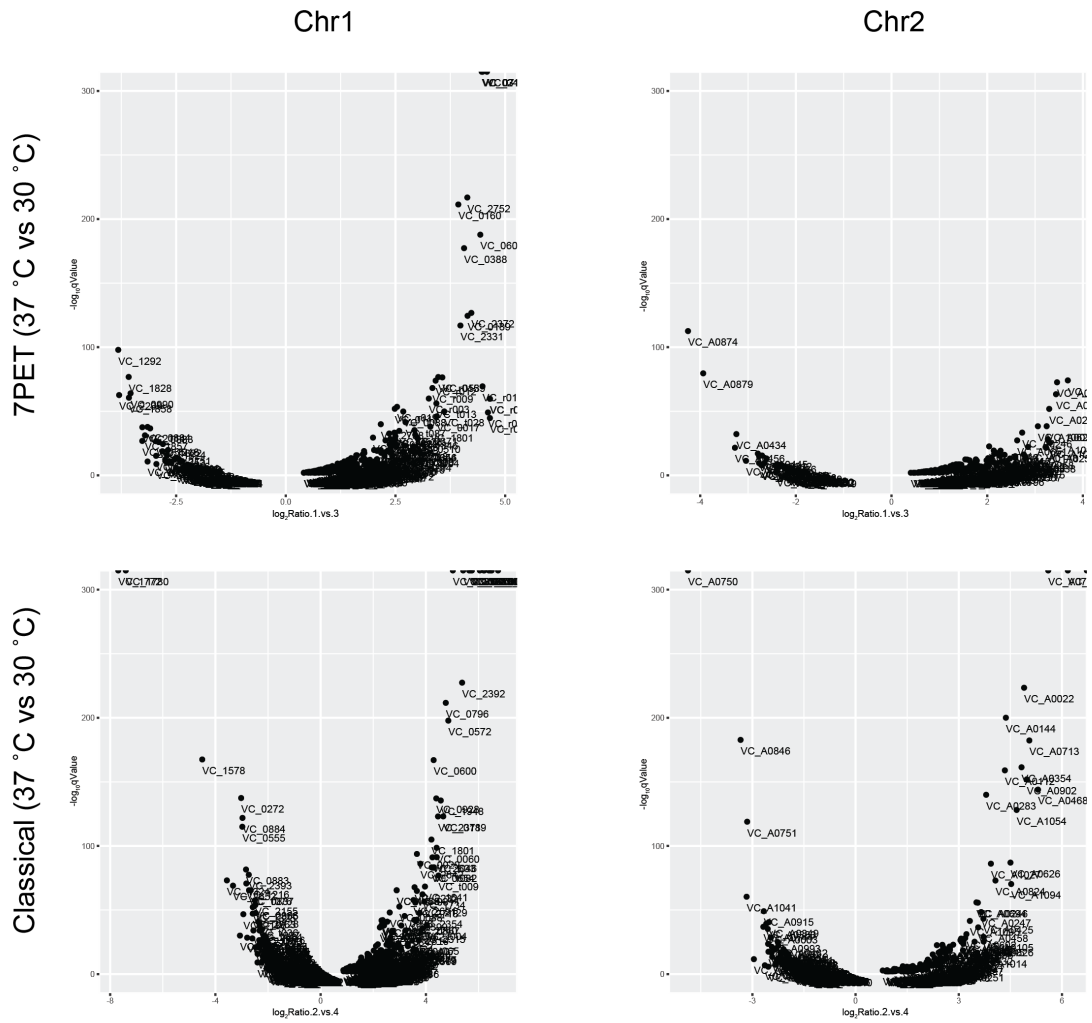


Figure 6.3 – Volcano plot comparing genes expressed at 37 and 30 ˚C in Classical and 7PET strains. Genes that were not detected in one strain or the other are not represented in these plots (log_2FC = infinite).

One of the most striking expression differences amongst these data was the effect of temperature on the expression of a subset of genes on VPI-2 in Classical *V. cholerae* (Figure 6.4). These genes encode sialidase and other enzymes required to catabolise and metabolise sialic acids (section 1.2.5). This difference in transcription is illustrated in Figure 6.4, and although at 30 ˚C there is a general increase in transcription of genes on VPI-2 in the Classical isolate, these sialic acid metabolism genes were specifically and dramatically up-regulated in comparison to their surrounding genes (Figure 6.4). It is important to note that neither VPI-2 nor the *nanH*/sialidase gene clusters are considered canonically to be part of the ToxR or ToxT regulons – therefore, further investigation is required to determine whether or not the up-regulation of virulence genes and sialidase genes at 30 ˚C in Classical *V. cholerae* are due to mechanisms or signals that are independent of one another.

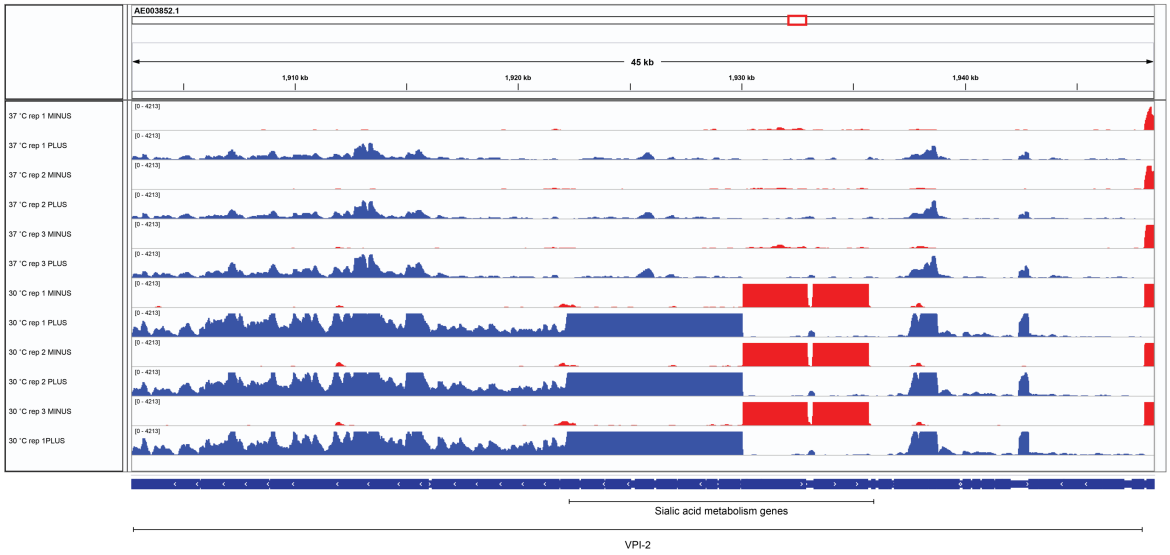


Figure 6.4 – Upregulation of sialic acid metabolisms genes in Classical *V. cholerae* at 30 ˚C relative to 37 ˚C. All lanes scaled identically.

There was also evidence that genes involved in *V. cholerae* virulence were up-regulated in Classical cultures at 30 ˚C relative to 37 ˚C under these *in vitro* conditions, exemplified by the *ctxAB* operon (illustrated in Figure 6.5). This is consistent with previous reports that have shown that culturing Classical *V. cholerae* at 30 ˚C induces virulence gene expression and the production both of CT and TCP, relative to culturing at 37 ˚C [91, 476–478]. The biological reason for this difference remains unclear. However, this has previously been shown not to be the case for 7PET bacteria, and this observation was recapitulated in our experiment – temperature variation did not cause statistically significant differential expression of virulence genes in cultures of 7PET *V. cholerae*.

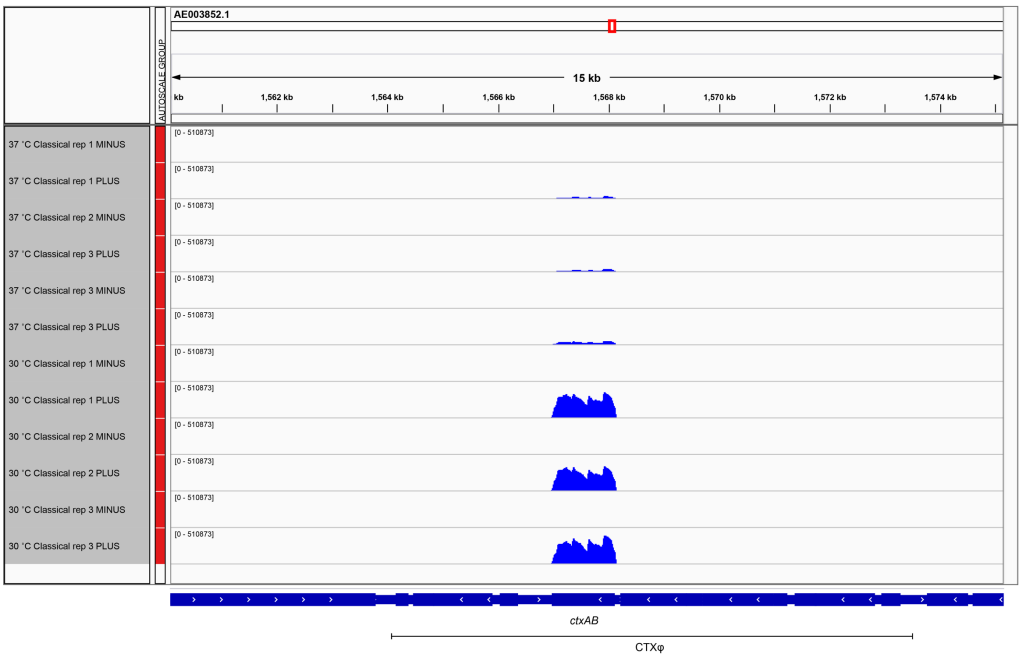


Figure 6.5 – Illustration of *ctxAB* transcript levels in Classical *V. cholerae* at 30 ˚C. The *ctxAB* operon is transcribed at very high levels at 30 ˚C in comparison to the levels at 37 ˚C. All lanes scaled identically.

6.3.5 – Selection of isolates for transcriptomic experiments

These initial experiments established that RNA-seq approaches could be used successfully with *V. cholerae* at WSI, and provided initial information about the practicalities of designing such experiments under CL3 conditions. Informed by these results, attention was then turned to extending these comparative transcriptomics methodologies to broader biological questions – specifically, investigating whether pandemic and non-pandemic lineages of *V. cholerae* O1 displayed different transcriptomic signatures when cultured under identical growth conditions.

Eight whole-genome-sequenced *V. cholerae* isolates were selected for comparative transcriptomic analysis. A number of factors were considered in choosing these isolates, including the complement of virulence genes and genomic islands which they harboured, their serogroup, the presence and absence of important promoter mutations, and their membership of key phylogenetic lineages. These factors were established in the course of the research described in Chapter 5. The strains are detailed in Table 6.6, and their phylogenetic position is indicated in Figure 6.6.

Internal ID	Other name	Serogroup	Details
MJD1402	NCTC 10256	O1	7PET. El Tor biotype. Toxigenic. Hong Kong, 1961
MJD1403	MJD474	O1	7PET. El Tor biotype. Toxigenic. UK traveller returning from Somalia, 2017
MJD1404	NCTC 10732	O1	Classical. Classical biotype. 1952
MJD1405	NCTC 5596	O1	Classical. Classical biotype. Pre-1939
MJD1406	A213	O1	Gulf Coast. El Tor biotype. Non-toxigenic. Georgia, 1984
MJD1407	A219	O1	Gulf Coast. El Tor biotype. Toxigenic. Georgia, 1986
MJD1408	MJD462	Non-O1	Related to MS-6. UK traveller returning from Thailand, 2017
MJD1409	NCTC 9422	O1	Non-7PET. El Tor biotype. Non-toxigenic. Pre-1955

Table 6.6 – Strains used in transcriptomic experiments. Each of these glycerol stocks were prepared from a single well-isolated colony taken from the original culture of each strain that was frozen at WSI. Isolation dates listed where known. Live isolates also listed in Table 2.1.

Two 7PET isolates were included, one from wave 1 (MJD1402), and a recent travel-associated isolate that phylogenetic analysis has shown to be a member of wave 3 (MJD1043). Wave 1 corresponds to the initial outbreak of the seventh cholera pandemic, and wave 3 to more recent pandemic cholera [234]. Both of these isolates are toxigenic, and harbour *ctxB3* and *ctxB7* alleles of *ctxB*, respectively. Phylogenetic analysis demonstrated that MJD1403 is a member of the same sub-lineage of 7PET as the isolates which caused the Yemeni cholera epidemic in 2017 – these isolates also harbour the *ctxB7* variant [309].

Two Classical isolates, MJD1404 and MJD1405, were included in this study. Both of these are toxigenic *V. cholerae* accessioned by NCTC, and harbour the *ctxB1* variant of *ctxB*, as expected for Classical *V. cholerae* [479]. These two isolates were included in the analysis because although MJD1404 is closely-related to the O395 reference sequence (Figure 5.10), MJD1405 harbours a wild-type P_{tcpPH} allele (Figure 5.20). Given the importance of this promoter mutation in virulence gene regulation (section 5.3.8.2), this second Classical isolate was included in the experiment.

Two *V. cholerae* O1 (A213, A219) that were members of the Gulf Coast lineage were included in these experiments. These were selected because they were non-toxigenic and toxigenic, respectively. Both originated from Georgia, USA, and were isolated in the 1980s. Their genome sequences were first reported in 2011 [234]. Like other members of the Gulf Coast lineage, these isolates harbour VPI-1 and VPI-2, but lack VSP-1 and VSP-2 (section 5.3.5). They are related to, but not identical to, strain 2740-80, the commonly-described representative of the Gulf Coast lineage [54, 189, 234]

Although this experiment was designed to explore whether there were differences in gene expression between 7PET and other *V. cholerae* O1, a single non-O1 isolate was included in this analysis. This non-toxigenic isolate was included because it is very closely related to MS-6, a toxigenic serogroup O1 isolate that harbours VPI-1, VPI-2, and VSP-1 [440]. MS-6 and MJD1408 are more closely related to Gulf Coast and MX-2 lineages, and to recent non-7PET Chinese *V. cholerae* O1 [396] than they are to 7PET. By including this isolate, similarities in gene expression to that of Gulf Coast isolates might be detectable which are independent of serogroup, though such candidate results would need to be validated.

The eighth strain included in the experiments was NCTC 9422, one of the non-pandemic El Tor *V. cholerae* O1 isolates characterised in Chapter 5 (section 5.3.7.2). This isolate is non-toxigenic and lacks all of the canonical *V. cholerae* pathogenicity islands, and was included to determine the transcriptional profile exhibited by a serogroup O1 biotype El Tor isolate that is more distantly related to 7PET than the Gulf Coast isolates.

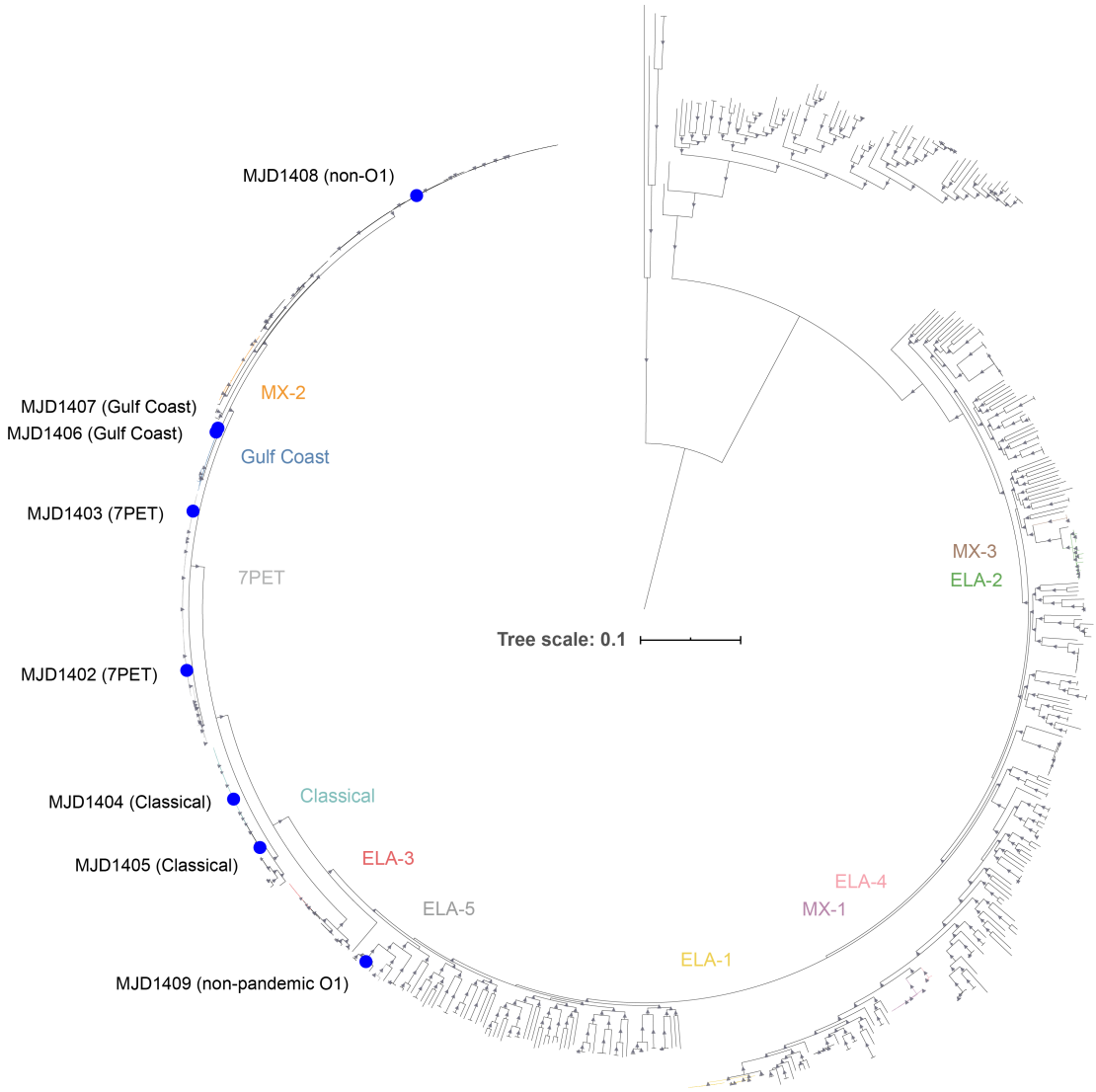


Figure 6.6 – Phylogenetic position of the live isolates chosen for this chapter research. Lineages are denoted as previously described [189]. The phylogeny presented here is the same as that used to generate Figure 5.1.

It is important to state that the chosen isolates do not represent the totality of *V. cholerae* diversity, and were chosen for the specific reasons outlined above. However, this multi-strain transcriptomic experiment was in itself a pilot experiment, designed to determine whether or not lineage-specific signatures of gene expression could be detected. This was also the maximum number of isolates that could be handled safely in one working day from a practical perspective. In order to assay gene expression across multiple strains within the constraints of working at CL3, cultures were harvested for RNA extraction after a defined period of time post-inoculation (3 hr), rather than by monitoring optical density for each culture and stopping the experiment for each strain once a defined OD_{600} value had been reached.

6.3.6 – Assaying differential gene expression in eight strains

Eight *V. cholerae* isolates were cultured at 37 ˚C in LB media, under identical conditions to those described in section 6.3.1. After 3 hr, cultures were harvested and RNA was subsequently isolated from snap-frozen cell pellets. All other conditions were kept identical to that of the pilot experiment (section 6.3.1). Sequencing data were analysed in the same way as was carried out in the pilot experiment, and genes that were statistically significantly differentially expressed were determined, pairwise, between each of the eight strains. These data have been summarised in Table 6.7.

		X							
		1402	**1403**	**1404**	**1405**	**1406**	**1407**	**1408**	**1409**
Y	**1402**	-	40/0	67/30	110/1077	60/11	55/14	61/55	409/1110
	1403		-	64/96	102/1119	49/72	42/140	1/237	369/1129
	1404			-	5/1058	43/54	47/77	44/118	259/1129
	1405				-	1045/82	1038/73	999/96	204/127
	1406					-	0/3	22/42	335/1110
	1407						-	24/40	272/1102
	1408							-	348/1077
	1409								-

Table 6.7 – Summary numbers for differentially-expressed genes across 8-way experiment. Number of genes up-regulated in isolate X relative to isolate Y / number of genes up-regulated in isolate Y compared to isolate X. Note that genes which were present in only one of these isolates were excluded from these aggregate numbers.

It was immediately apparent that MJD1405 and MJD1409 exhibited drastically different gene expression profiles to all other isolates - over 1,000 genes were expressed in these isolates relative to all other isolates in the experiment (Table 6.7). Moreover, gene expression in these two isolates was much more similar between one another than between these and other isolates, (Table 6.7). To investigate this using a complementary approach, a principal component analysis (PCA) of normalised transcripts for all 24 samples was performed to summarise the variation across all of the samples (Figure 6.7). Principal component 1 (PC1), explaining 57.3% of the variation amongst these 24 isolates, clearly separated both MJD1405 and MJD1409 from the rest of the experiment, and also showed that the three replicates for each of these isolates were dissimilar to one another (Figure 6.7). The variability within these replicates is unlikely to be explained by differences in growth stage – all cultures were harvested at the same time

post-inoculation, and OD_{600} values for each replicate were similar at the time of harvesting. It was therefore decided not to compare the data from these two strains to the others collected in this experiment.

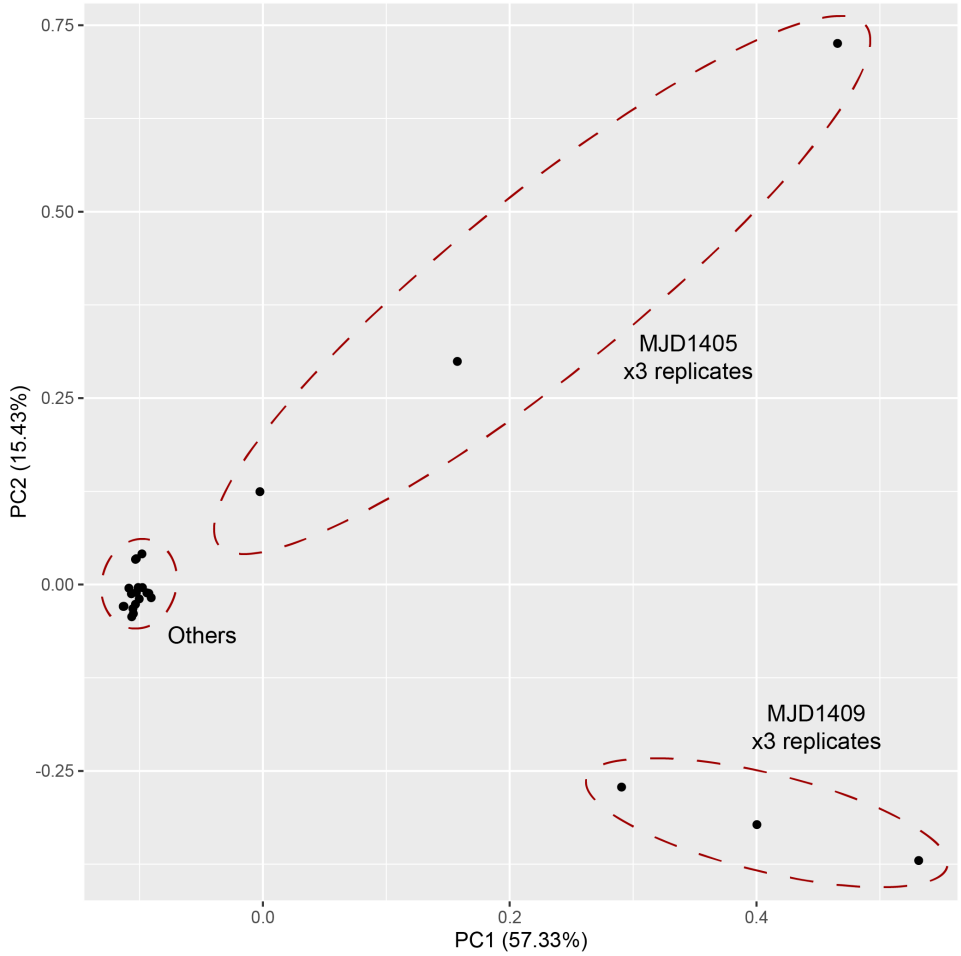


Figure 6.7 – PCA comparing all of the 24 sequenced samples in this experiment. The biological replicates of MJD1405 and MJD1409 were substantially different to the other strains in the experiment and to one another, separated by PC1. Ellipses added manually for illustrative purposes (Adobe Illustrator).

PCA was then repeated after excluding replicates of MJD1405 and MJD1409, on eighteen samples in total (Figure 6.8). In this analysis, PC1 (explaining 26.3% of the variance amongst these replicates) clearly separated Classical isolate MJD1404 from the rest of the samples. However, there appeared to be limited separation of samples along PC2; replicates of different strains were mixed amongst one another, and the three replicates of each strain did not cluster with one another in every instance. This was a surprising result, as it had been hypothesised that isolates would cluster on the basis of their genotype or membership of genetic lineage. This may reflect variation within the experiment, or that each culture had reached a slightly

different growth phase stage at the time of harvesting. This variability meant that we were conservative in selecting genes for subsequent consideration or discussion in the absence of confirmatory experiments.

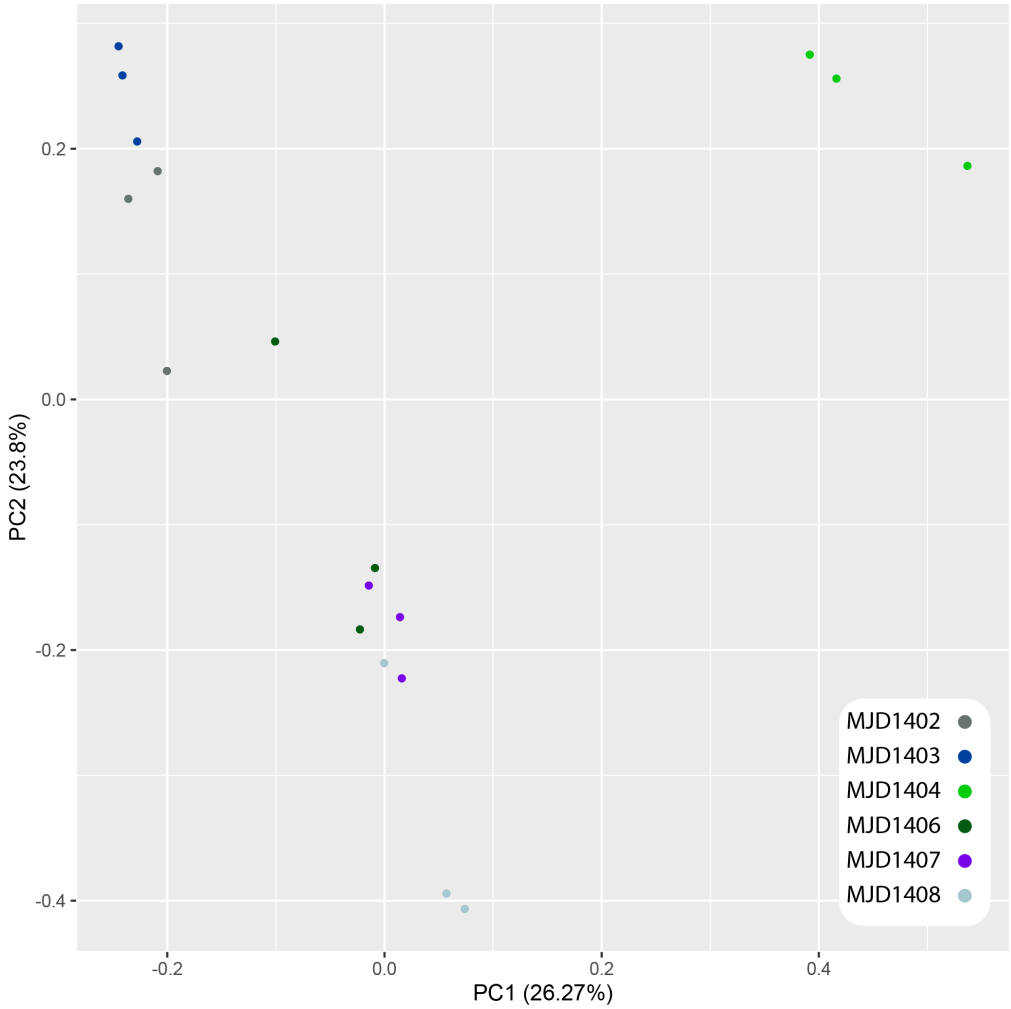


Figure 6.8 – PCA of the 18 samples remaining after the exclusion of MJD1405 and MJD1409. The biological replicates of MJD1404, the Classical isolate remaining in the experiment, were clearly separated from the other samples in the experiment by PC1 and PC2. However, the remaining samples failed to resolve into clearly-separated groups based on their genotype.

It was decided to interrogate these data to identify potentially novel biological insights. For example, to our knowledge, this experiment is the first to compare the transcriptome of Gulf Coast isolates to those of either Classical or 7PET bacteria. Amongst the two Gulf Coast strains, just three genes were significantly differentially expressed (Table 6.7). These were *VC_1865*, *VC_A0540*, and one putative CDS of unknown function. Of these, only the product

of *VC_A0540* has a known or predicted function (formate transporter). This lack of variability is consistent with one of the aims of this experiment – namely, to determine whether strains within a *V. cholerae* lineage have transcriptional profiles more similar to the profiles of other lineages.

We interrogated the genes differentially expressed amongst Gulf Coast and 7PET isolates (Table 6.7), and immediately found it to be striking that multiple components of T6SS were up-regulated in Gulf Coast isolates in comparison to 7PET isolates. The macromolecular T6SS is an apparatus that enables *V. cholerae* to prey upon bacteria and eukaryotes, by secreting and injecting effector proteins into target cells (*e.g.*, [480, 481]). Three such systems are usually encoded by three gene clusters in *V. cholerae*, and among the genes involved in producing T6SS is *hcp*, which encodes the Hcp protein (*VC_1415*, *VC_A0017*), and is considered to be a 'hallmark' of T6SS function [482]. Hcp is thought to form hexameric rings that constitute the tube through which T6SS secrete effectors [483]. These *hcp* genes, and other genes encoding structural components of *V. cholerae* T6SS, were substantially differentially expressed in these isolates, exemplified in Figure 6.9. Examining transcription profiles across all replicates and strains also showed that MJD1408 exhibited a similar *hcp* transcription profile to Gulf Coast isolates (to which it is most closely related phylogenetically, Figure 6.6).

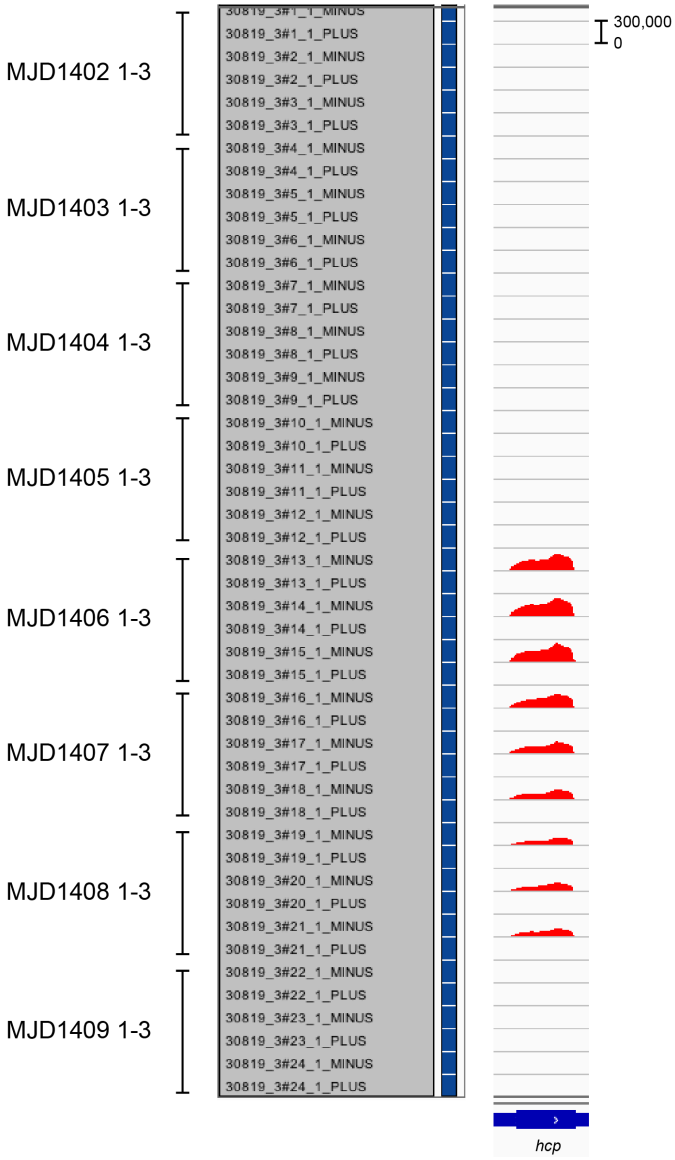


Figure 6.9 – The *hcp* gene (*VC_A0017*) is upregulated in Gulf Coast and a related strain of *V. cholerae* (MJD1408) relative to all other strains in this experiment. All tracks scaled identically.

6.4 – Discussion

The design of these comparative experiments, and these initial data, lay the groundwork for future analyses that should begin to describe lineage-specific as well as strain-specific differences in gene expression amongst *V. cholerae*. These data have shown that genes which are highly over- or under-expressed can be identified, and have also highlighted the homogeneity of expression profiles within lineages and other closely-related *V. cholerae*. Characterising aspects of gene regulation that are hallmarks of a *V. cholerae* lineage should add to our understanding of the biological differences between pandemic and non-pandemic bacterial lineages. It is also notable that these data recapitulated some, but not all, of the previously-published differences in gene expression between Classical and 7PET *V. cholerae* [470]. The genes that were detected in both studies were highly differential-expressed in both studies, and this suggests that in spite of the differences in growth conditions between these experiments, any fundamental and dramatic lineage-specific gene expression patterns can still be detected (such as the hyperactivation of the ToxR regulon in Classical isolates grown *in vitro*). This is reinforced by the fact that the strains used in both experiments were different, though members of the same phylogenetic lineages. This suggests that this approach of comparing gene expression across lineages has merit, and although variation between experiments might influence expression patterns, dramatically-different expression profiles may still be detectable.

However, these data also highlight the risks inherent in working with diverse bacteria under defined *in vitro* conditions. Variation in growth rates, auxotrophies, and the presence of mutations that lower fitness *in vitro* are all factors which could introduce confounding variation into experiments, including the large experiment described in this chapter. Such variations might also explain why replicate cultures of MJD1405 and MJD1409 did not resemble one another in PCA analyses (Figure 6.7). Nevertheless, these data show that observations such as differential expression of virulence genes and T6SS can be made in such an experiment, provided the magnitude of the signal in question is large enough to be detected and to be trustworthy. Additional experiments such as qRT-PCR and the construction of targeted mutants will still be required to validate results from transcriptomic experiments. It is also evident from these data that using this approach, dramatic differences in gene expression can be detected between strains and across growth conditions (Figures 6.5, 6.9). This lends itself to the development of genetic tools, such as fusions between differentially-regulated promoters

(*e.g.*, P_{hcp}, P_{ctxAB}, P_{nanH}) and reporter genes such as *gfp* and *lux*, with which to identify *cis* and *trans*-factors involved in these differences in gene expression. Using the large collection of genome sequences now available to us, the distribution of any such factors, or variants of factors, can be easily determined across the sequenced *V. cholerae* species, such as was carried out in sections 3.4.9, 4.3.11, 5.3.3, and 5.3.8.

The stark differences in T6SS gene expression in multiple *V. cholerae* were particularly intriguing because it has been shown that the 2740-80 Gulf Coast strain of *V. cholerae* O1 constitutively expresses T6SS *in vitro*, whereas in 7PET bacteria, suitable inductive signals must be present for T6SS to be expressed (*e.g.*, [50, 484]). Similarly, the O37 isolate V52 expresses T6SS constitutively *in vitro*, and this strain has been used to study the biology of T6SS in *V. cholerae* [485, 486]. Crucially, although 2740-80 has been shown to express T6SS constitutively, neither of the Gulf Coast strains used in this analysis are identical to 2740-80 [234]. This strongly suggests that the constitutive expression of T6SS is a feature of the Gulf Coast lineage. Moreover, MJD1408 is closely related to the Gulf Coast lineage (Figure 6.6), suggesting that this T6SS expression phenotype might be common to other related lineages within *V. cholerae*. Testing of additional isolates, including those that belong to MX-2 or resemble the Chinese *V. cholerae* O1 that are related to the Gulf Coast lineage (section 5.3.1), may be warranted.

The VasH protein [487], the sigma factor RpoN [485], exposure to chitin, and quorum sensing [50, 484] are all factors which regulate the expression of T6SS in *V. cholerae*. VasH is different between N16961 and V52, and ectopic expression of $VasH_{V52}$ can drive aspects of T6SS expression in N16961, but is not sufficient to activate full T6SS killing activity *in vitro* [487]. The VasH protein sequence has also been shown to be variable across non-O1/O139 *V. cholerae* [488, 489]. The constitutive expression of T6SS is an area of current research and although the mechanisms by which this expression is regulated constitutively in V52 remain unclear, the activity of the WigKR two-component system has recently been implicated in governing T6SS regulation in V52 [490]. In future work, it will be necessary to determine the variation of *vasH* amongst these isolates and across the *V. cholerae* phylogeny, and to determine whether introducing variation in this or other regulators can elicit a constitutive expression phenotype in 7PET or Classical isolates (or, indeed, if abolishing these regulators' activity in Gulf Coast isolates can confer a 7PET-like T6SS expression phenotype *in vitro*).

Future experiments such as these should consider the utilisation of alternative growth conditions, to characterise further any gene expression differences that are lineage-specific. For instance, LB liquid media was used in all of these experiments. It would be worthwhile using minimal media supplemented with relevant carbon sources (*e.g.*, glucose, chitin, etc), both to determine the ability of chosen strains to grow on these media, and to assess whether these nutritional stresses provoke differences in the transcriptome of these isolates and lineages.

It will also be important to repeat these analyses using a panel of strains that includes the reference *V. cholerae* strains that are used by molecular biologists, to ascertain with confidence whether the previous work published on those reference strains describes the transcriptional behaviour of those strains, or of lineages more generally. Similarly, it is important to recognise that the gene expression data calculated in these experiments were generated by mapping to the N16961 reference genome. Differences in the expression of genes present in other isolates that are absent from N16961 will not be detectable using this approach. Producing closed genome assemblies for each strain being assayed will be an important next step in the optimisation of these large-scale transcriptomic assays.

Chapter 7

Summary and future directions

7.1 – Summary of thesis findings

In this thesis, I have used a combination of *in silico* and *in vitro* approaches to increase our understanding of the differences between pandemic and non-pandemic lineages of *Vibrio cholerae*. The initial thesis aims (section 1.5) were to study the evolution of 7PET longitudinally and across a large geography, and to characterise the non-7PET *V. cholerae* present in a country during a pandemic, alongside an external introduction of 7PET. From that baseline, the thesis then set out to investigate further the diversity of non-epidemic and non-7PET *V. cholerae*, and the functional differences between pandemic and non-pandemic lineages of *V. cholerae* O1. This involved the detailed investigation of single fully-assembled genome sequences for important bacterial isolates, the collective examination of large numbers of genomes, and the design and execution of transcriptomic experiments.

In Chapter 3, I presented a study of Argentinian *V. cholerae* which has increased our understanding of how 7PET evolves, and does not evolve, following its introduction into a country and population that was naïve to pandemic cholera caused by 7PET. These findings have also been actionable, contributing to changes in public health policies in Argentina (Chapter 3). Very minimal change in 7PET was observed, in terms of SNV accumulation, gene gain/loss, and recombination, though important conclusions were drawn from the observed variation in Inaba/Ogawa serotype and genotype, and the potentially-misleading conclusions that might be drawn from looking at phenotypic serotype variation alone. The clonality of 7PET contrasted dramatically with that of the non-7PET *V. cholerae* from Argentina that were included in this thesis research. Many of these bacteria were isolated from suspected cholera cases, and the roles of virulence determinants other than CTXφ, such as T3SS, remains to be investigated. These non-7PET *V. cholerae* present in Argentina throughout the cholera epidemic are proposed to represent the population of *V. cholerae* that are truly endemic to Argentina, which may or may not cause symptoms that resemble clinical cholera.

In Chapter 4, I illustrated the insights that can be gleaned from the study of high-quality closed genome assemblies. I presented novel and unique aspects of the genome biology of recently-isolated *V. cholerae* O139, such as the presence of multiple *ctxB* alleles in the same genome, and cross-chromosomal duplications of VSP-1. I also described a genome assembly for non-toxigenic *V. cholerae* O139 which occupies an unusual position in the *V. cholerae* phylogeny. Furthermore, as part of developing methods for work with *V. cholerae* at CL3, I assembled a

closed genome assembly for NCTC 30, which we believe to be the world's oldest publicly-available non-O1 *V. cholerae*. Using this genome sequence alongside complementary experimental methods, I confirmed the presence of a functional β-lactamase in this isolate (which pre-dates the introduction of antimicrobials as therapeutics), as well as the presence of a T3SS-2β element, and provided plausible explanations for the absence of flagella in this strain.

Following this detailed characterisation of closed *V. cholerae* genome sequences, I proceeded to analyse specific, relevant genotypes and phenotypes across a set of 646 highly-diverse *V. cholerae* (Chapter 5). These included virulence determinants, antimicrobial resistance genes, and determinants of *V. cholerae* O1 biotypes, which are relevant to clinicians, to molecular microbiologists, and to those who study *V. cholerae* genomics. These data reinforced a number of previous findings, including the fact that very few plasmid replicons were detectable amongst diverse *V. cholerae* as well as within 7PET, and that T3SS virulence determinants are increasingly common amongst these diverse isolates, both those of clinical and environmental origin. This work also led to the first sequence of an historical IncA/C2 plasmid devoid of antimicrobial or heavy-metal resistance genes. This backbone may represent an "ancestral" IncA/C2 plasmid, and future work to perform phylogenetics of the many IncA/C plasmids being sequenced from *V. cholerae* and other species would be worthwhile.

These data also highlighted inconsistencies amongst previous observations that had been considered to be dogma in *V. cholerae* biology. For instance, I identified numerous examples of *V. cholerae* pathogenicity islands present in non-pandemic isolates and lineages. Canonically, these elements have been described to be unique to epidemic and pandemic lineages, and this observation emphasises the fact that as we sequence additional members of this species, our definition of these truisms and our understanding of whether a mobile element is a 'marker' for a lineage will need to be refined. Similarly, by studying historical El Tor biotype *V. cholerae* O1, I have shown that the phenotypes associated with the classical biotype, and the mutations leading to these phenotypes, collectively only describe the Classical lineage – the El Tor biotype is broadly distributed amongst diverse *V. cholerae*, irrespective of serogroup. These results highlighted and emphasised that biotyping reactions are of limited utility in modern studies of *V. cholerae*, suggesting that the microbiological resources invested in biotyping might be better employed in other ways. However, this is wholly understandable in the context of how our understanding of *V. cholerae* and of cholera has developed and

evolved since the bacterium was first observed by Pacini. As our understanding has evolved, and as technologies with which to study *V. cholerae* have improved from biochemistry to whole-genome sequencing, the resolution at which we understand this pathogen has grown. These refinements to our understanding mean that current cholera control strategies, such as those of the GTFCC, now have the opportunity to effect meaningful change to global health.

In Chapter 6, I describe transcriptomic experiments designed and executed in order to characterise lineage-specific differences in gene expression amongst *V. cholerae* O1. These were pilot experiments, designed both to optimise transcriptomic methodologies for use in CL3, and to discern robust differences in gene expression across *V. cholerae* lineages. These experiments have provided novel results as well as recapitulating the results from previous studies, which used different bacterial strains. These include apparent lineage-specific differences in T6SS regulation, and lineage-specific temperature-mediated differential expression of virulence and pathogenicity genes. The work presented in this chapter also highlights the difficulties in working with heterogenous bacteria, and the need to balance optimal experimental design with realistic practical considerations.

7.2 – Future directions

Cholera is an ancient disease, which we have failed to control to date, although efforts to do so are ongoing. It is clear from this thesis research, as well as from a growing body of literature, that taking a broad approach to sequencing *V. cholerae* – *i.e.*, expanding focus from the exclusive sequencing of epidemics – is generating a wealth of genomic diversity for future analysis. Although the data in this thesis has advanced our understanding of the *V. cholerae* species as discussed above and in previous chapters, they also highlight open questions in this research area, and lay the groundwork by which to investigate these questions. In future projects, sequencing and functional research must not only focus on members of 7PET, but also on environmentally- and clinically-isolated *V. cholerae* irrespective of their serogroup and toxigenicity. Efforts should be invested in carrying out comprehensive studies of the pangenome of this species as a whole, alongside functional assays to determine the transcriptomic, biochemical, and metabolic differences amongst diverse *V. cholerae*. Particular attention should be paid to the clade of diverse isolates highlighted throughout this thesis (Chapters 4, 5). It is only by adopting such an holistic approach, aiming to capture all *V. cholerae* including those that do not appear to cause disease, that we will amass sufficient

genomic data and biological material with which to explore the question of what enables pandemic lineages to cause pandemics.

In future work, it will be important to determine the significance of the lack of genomic variation amongst 7PET. It is unknown whether this lack of variation be explained mechanistically, and whether this is due to a reduced ability, or reduced opportunities, to participate in HGT. It is also unknown whether the pandemic lineage has fixed advantageous phenotypes by reducing genomic variation. This is a particularly relevant consideration because the reasons why 7PET can spread rapidly and to cause global pandemics are still uncertain. Further functional research into this area will be required to explain this observation. Similarly, the apparent lack of plasmids amongst *V. cholerae*, even amongst diverse isolates (Chapter 5), should also be studied in more detail. It is unknown why IncA/C plasmids are one of the very few plasmids that this species appears to be able to maintain. It will also be important to pursue this research avenue without assuming that databases of plasmid replicons contain all known *Vibrio* plasmids – it might be that genomic analyses fail to identify *V. cholerae* plasmids because the Plasmidfinder database simply lacks the replicon sequences for plasmids found in this species. Coupling genome sequencing to wet-lab bacteriology, and functional studies of plasmid dynamics and diversity, will expand our knowledge of HGT mechanisms and potential within this species.

The role of T3SS in disease mediated by *V. cholerae* is also a research area with great potential. Although we found in this thesis that many clinically-isolated *V. cholerae* encoded T3SS in the absence of other canonical virulence factors, causal links between these systems and pathology in patients have not yet been proven. Considerable opportunity exists to exploit the genome sequences generated in this thesis to characterise in detail the diversity of T3SS systems *in silico*, as well as to avail of having access to live isolates of multiple T3SS^{+} *V. cholerae* to study the regulation of these systems, their mobility amongst different *Vibrio* spp., and their functionality in models of pathogenesis.

The potential to discover new aspects of *V. cholerae* biology using this holistic approach – studying *V. cholerae* beyond 7PET – is exemplified by the transcriptomic studies set out in this thesis (Chapter 6). However, it is important to bear in mind that although it is logical and plausible that transcriptomic responses to stresses and signals would be more similar amongst members of the same lineage than across lineages, this remains an hypothesis and must be

tested *in vitro* and *in vivo*. The work in this thesis has set the stage for subsequent experiments to be refined and optimised, to test these hypotheses, and to expand our understanding of the *V. cholerae* species beyond that of 7PET and Classical lineages.

References

1. **Kaper JB, Morris JG, Levine MM**. Cholera. *Clinical Microbiology Reviews* 1995;8:48–86.

2. **Clemens JD, Nair GB, Ahmed T, Qadri F, Holmgren J**. Cholera. *The Lancet* 2017;390:1539–1549.

3. **Centers for Disease Control and Prevention (CDC)**. General information | Cholera | CDC. https://www.cdc.gov/cholera/general/index.html (2018, accessed 14 September 2018).

4. **Ali M, Nelson AR, Lopez AL, Sack DA**. Updated global burden of cholera in endemic countries. *PLoS Neglected Tropical Diseases* 2015;9:e0003832.

5. **Ali M, Lopez AL, You YA, Kim YE, Sah B, *et al.*** The global burden of cholera. *Bulletin of the World Health Organization* 2012;90:209-218A.

6. **World Health Organization**. Cholera. *Cholera*. https://www.who.int/news-room/fact-sheets/detail/cholera (2019, accessed 21 June 2020).

7. **Kim J-H, Mogasale V, Burgess C, Wierzba TF**. Impact of oral cholera vaccines in cholera-endemic countries: A mathematical modeling study. *Vaccine* 2016;34:2113–2120.

8. **Talavera A, Pérez EM**. Is cholera disease associated with poverty? *Journal of Infection in Developing Countries* 2009;3:408–411.

9. **World Health Organization**. Ending cholera: A global roadmap to 2030. http://www.who.int/cholera/publications/global-roadmap.pdf (2017, accessed 14 September 2018).

10. **Centers for Disease Control and Prevention (CDC)**. Rehydration therapy | Treatment | Cholera | CDC. https://www.cdc.gov/cholera/treatment/rehydration-therapy.html (2018, accessed 23 March 2020).

11. **Centers for Disease Control and Prevention (CDC)**. Cholera - *Vibrio cholerae* infection | Recommendations for the use of antibiotics for the treatment of cholera | Cholera | CDC. https://www.cdc.gov/cholera/treatment/antibiotic-treatment.html (2015, accessed 18 March 2018).

12. **Lindenbaum J, Greenough WB, Islam MR**. Antibiotic therapy of cholera in children. *Bulletin of the World Health Organization* 1967;37:529–538.

13. **World Health Organization**. *Cholera outbreak | Assessing the outbreak response and improving preparedness*. World Health Organization; 2004.

14. **Pan American Health Organisation (PAHO)**. *Recommendations for clinical management of cholera*. Washington (DC); 2010.

15. **Nelson EJ, Klarman M**. *Cholera outbreak training and shigellosis program*. 2nd ed. icddr,b; 2018.

16. **Médecins Sans Frontières**. *Cholera guidelines*. 2nd ed. France; 2004.

17. **World Health Organization**. Cholera vaccines: WHO position paper - August 2017. *Weekly Epidemiological Record* 2017;92:477–500.

18. **Levine MM, Herrington D, Losonsky G, Tall B, Kaper JB, *et al.*** Safety, imunogenicity, and efficacy of recombinant live oral cholera vaccines, CVD 103 and CVD 103-HgR. *The Lancet* 1988;332:467–470.

19. **Ketley JM, Michalski J, Galen J, Levine MM, Kaper JB**. Construction of genetically marked *Vibrio cholerae* O1 vaccine strains. *FEMS Microbiology Letters* 1993;111:15–21.

20. **Shaikh H, Lynch J, Kim J, Excler J-L**. Current and future cholera vaccines. *Vaccine* 2020;38:A118–A126.

21. **Qadri F, Azad AK, Flora MS, Khan AI, Islam MT, *et al.*** Emergency deployment of oral cholera vaccine for the Rohingya in Bangladesh. *The Lancet* 2018;391:1877–1879.

22. **Hubbard TP, Billings G, Dörr T, Sit B, Warr AR, *et al.*** A live vaccine rapidly protects against cholera in an infant rabbit model. *Science Translational Medicine* 2018;10: eaap8423.

23. **Fakoya B, Sit B, Waldor MK**. Transient intestinal colonization by a live-attenuated oral cholera vaccine induces protective immune responses in streptomycin-treated mice. *Journal of Bacteriology*. Epub ahead of print 15 June 2020. DOI: 10.1128/JB.00232-20.

24. **Sit B, Zhang T, Fakoya B, Akter A, Biswas R, *et al.*** Oral immunization with a probiotic cholera vaccine induces broad protective immunity against *Vibrio cholerae* colonization and disease in mice. *PLoS Neglected Tropical Diseases* 2019;13:e0007417.

25. **Piarroux R, Barrais R, Faucher B, Haus R, Piarroux M, *et al.*** Understanding the cholera epidemic, Haiti. *Emerging Infectious Diseases* 2011;17:1161–1168.

26. **Chin C-S, Sorenson J, Harris JB, Robins WP, Charles RC, *et al.*** The origin of the Haitian cholera outbreak strain. *New England Journal of Medicine* 2011;364:33–42.

27. **Hendriksen RS, Price LB, Schupp JM, Gillece JD, Kaas RS, *et al.*** Population genetics of *Vibrio cholerae* from Nepal in 2010: Evidence on the origin of the Haitian outbreak. *mBio* 2011;2:e00157-11.

28. **Camacho A, Bouhenia M, Alyusfi R, Alkohlani A, Naji MAM, *et al.*** Cholera epidemic in Yemen, 2016–18: An analysis of surveillance data. *The Lancet Global Health* 2018;6:e680–e690.

29. **Colwell RR**. Global climate and infectious disease: The cholera paradigm. *Science* 1996;274:2025–2031.

30. **Snow J**. *On the mode of communication of cholera*. 2nd ed. England: John Churchill; 1855.

31. **Bynum W**. In retrospect: On the mode of communication of cholera. *Nature* 2013;495:169–170.

32. **Centers for Disease Control and Prevention (CDC)**. Diagnosis and detection | Cholera | CDC. https://www.cdc.gov/cholera/diagnosis.html (2018, accessed 25 June 2020).

33. **Howard-Jones N**. Robert Koch and the cholera vibrio: a centenary. *British Medical Journal (Clinical Research Edition)* 1984;288:379–381.

34. **Pacini F**. Osservazioni microscopiche e deduzioni patologiche sul cholera asiatico. *Gazzetta Medica Italiana* 1854;397–405.

35. **Skerman VBD, McGowan V, Sneath PHA**. Approved lists of bacterial names. *International Journal of Systematic and Evolutionary Microbiology,* 1980;30:225–420.

36. **Pollitzer R, Swaroop S, Burrows W, World Health Organization**. *Cholera*. World Health Organization. https://apps.who.int/iris/handle/10665/41711 (1959, accessed 2 October 2019).

37. **Shimada T, Arakawa E, Itoh K, Okitsu T, Matsushima A, *et al.*** Extended serotyping scheme for *Vibrio cholerae*. *Current Microbiology* 1994;28:175–178.

38. **Colwell RR**. Polyphasic taxonomy of the genus *Vibrio*: Numerical taxonomy of *Vibrio cholerae*, *Vibrio parahaemolyticus*, and related *Vibrio* species. *Journal of Bacteriology* 1970;104:410–433.

39. **Davis GH, Park RW**. A taxonomic study of certain bacteria currently classified as *Vibrio* species. *Journal of General Microbiology* 1962;27:101–119.

40. **Furniss AL, Lee JV, Donovan TJ**. *The Vibrios*. London: His Majesty's Stationery Office; 1978.

41. **Gardner AD, Venkatraman KV**. The antigens of the cholera group of Vibrios. *Journal of Hygiene (London)* 1935;35:262–282.

42. **Colwell RR, Kaper J, Joseph SW**. *Vibrio cholerae*, *Vibrio parahaemolyticus*, and other vibrios: occurrence and distribution in Chesapeake Bay. *Science* 1977;198:394–396.

43. **del Refugio Castañeda Chávez M, Sedas VP, Orrantia Borunda E, Reynoso FL**. Influence of water temperature and salinity on seasonal occurrences of *Vibrio cholerae* and enteric bacteria in oyster-producing areas of Veracruz, México. *Marine Pollution Bulletin* 2005;50:1641–1648.

44. **Almagro-Moreno S, Taylor RK**. Cholera: Environmental reservoirs and impact on disease transmission. *Microbiology Spectrum* 2013;1:OH-0003-2012.

45. **Huq A, West PA, Small EB, Huq MI, Colwell RR**. Influence of water temperature, salinity, and pH on survival and growth of toxigenic *Vibrio cholerae* serovar 01 associated with live copepods in laboratory microcosms. *Applied and Environmental Microbiology* 1984;48:420–424.

46. **Singleton FL, Attwell R, Jangi S, Colwell RR**. Effects of temperature and salinity on *Vibrio cholerae* growth. *Applied and Environmental Microbiology* 1982;44:1047–1058.

47. **Meibom KL, Li XB, Nielsen AT, Wu C-Y, Roseman S, *et al.*** The *Vibrio cholerae* chitin utilization program. *Proceedings of the National Academy of Sciences of the United States of America* 2004;101:2524–2529.

48. **Meibom KL, Blokesch M, Dolganov NA, Wu C-Y, Schoolnik GK**. Chitin induces natural competence in *Vibrio cholerae*. *Science* 2005;310:1824–1827.

49. **Rowe-Magnus DA, Guerout A-M, Biskri L, Bouige P, Mazel D**. Comparative analysis of superintegrons: Engineering extensive genetic diversity in the Vibrionaceae. *Genome Research* 2003;13:428–442.

50. **Borgeaud S, Metzger LC, Scrignari T, Blokesch M**. The type VI secretion system of *Vibrio cholerae* fosters horizontal gene transfer. *Science* 2015;347:63–67.

51. **Matthey N, Stutzmann S, Stoudmann C, Guex N, Iseli C, *et al.*** Neighbor predation linked to natural competence fosters the transfer of large genomic regions in *Vibrio cholerae*. *eLife* 2019;8:e48212.

52. **Waldor MK, Mekalanos JJ**. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* 1996;272:1910–1914.

53. **Carraro N, Matteau D, Luo P, Rodrigue S, Burrus V**. The master activator of IncA/C conjugative plasmids stimulates genomic islands and multidrug resistance dissemination. *PLoS Genetics* 2014;10:pgen.1004714.

54. **Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, *et al.*** Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 2009;106:15442–15447.

55. **Karaolis DKR, Johnson JA, Bailey CC, Boedeker EC, Kaper JB, *et al.*** A *Vibrio cholerae* pathogenicity island associated with epidemic and pandemic strains. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95:3134–3139.

56. **Wozniak RAF, Fouts DE, Spagnoletti M, Colombo MM, Ceccarelli D, *et al.*** Comparative ICE genomics: Insights into the evolution of the SXT/R391 family of ICEs. *PLoS Genetics* 2009;5:e1000786.

57. **Trucksis M, Michalski J, Deng YK, Kaper JB**. The *Vibrio cholerae* genome contains two unique circular chromosomes. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95:14464–14469.

58. **Val M-E, Marbouty M, de Lemos Martins F, Kennedy SP, Kemble H, *et al.*** A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. *Science Advances* 2016;2:sciadv.1501914.

59. **Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, *et al.*** DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* 2000;406:477–483.

60. **Ramachandran R, Ciaccia PN, Filsuf TA, Jha JK, Chattoraj DK**. Chromosome 1 licenses chromosome 2 replication in *Vibrio cholerae* by doubling the *crtS* gene dosage. *PLoS Genetics* 2018;14:e1007426.

61. **Sozhamannan S, Waldminghaus T**. Exception to the exception rule: synthetic and naturally occurring single chromosome *Vibrio cholerae*. *Environmental Microbiology* Epub ahead of print 2020. DOI: 10.1111/1462-2920.15002.

62. **Johnson SL, Khiani A, Bishop-Lilly KA, Chapman C, Patel M, *et al.*** Complete genome assemblies for two single-chromosome *Vibrio cholerae* isolates, strains 1154-74 (serogroup O49) and 10432-62 (serogroup O27). *Genome Announcements* 2015;3:genomeA.00462-15.

63. **Bruhn M, Schindler D, Kemter FS, Wiley MR, Chase K, *et al.*** Functionality of two origins of replication in *Vibrio cholerae* strains with a single chromosome. *Frontiers in Microbiology* 2018;9:fmicb.2018.02932.

64. **Das B, Chattoraj DK**. Commentary: Functionality of two origins of replication in *Vibrio cholerae* strains with a single chromosome. *Frontiers in Microbiology* 2019;10:fmicb.2019.01314.

65. **Mazel D, Dychinco B, Webb VA, Davies J**. A distinctive class of integron in the *Vibrio cholerae* genome. *Science* 1998;280:605–608.

66. **Kirkup BC, Chang L, Chang S, Gevers D, Polz MF**. *Vibrio* chromosomes share common history. *BMC Microbiology* 2010;10:137.

67. **De SN**. Enterotoxicity of bacteria-free culture-filtrate of *Vibrio cholerae*. *Nature* 1959;183:1533–1534.

68. **Kaper JB, Moseley SL, Falkow S**. Molecular characterization of environmental and nontoxigenic strains of *Vibrio cholerae*. *Infection and Immunity* 1981;32:661–667.

69. **Moseley SL, Falkow S**. Nucleotide sequence homology between the heat-labile enterotoxin gene of *Escherichia coli* and *Vibrio cholerae* deoxyribonucleic acid. *Journal of Bacteriology* 1980;144:444–446.

70. **Sporecke I, Castro D, Mekalanos JJ**. Genetic mapping of *Vibrio cholerae* enterotoxin structural genes. *Journal of Bacteriology* 1984;157:253–261.

71. **Mekalanos JJ**. Duplication and amplification of toxin genes in *Vibrio cholerae*. *Cell* 1983;35:253–263.

72. **Zhang R-G, Scott DL, Westbrook ML, Nance S, Spangler BD, *et al.*** The three-dimensional crystal structure of cholera toxin. *Journal of Molecular Biology* 1995;251:563–573.

73. **Hardy SJ, Holmgren J, Johansson S, Sanchez J, Hirst TR**. Coordinated assembly of multisubunit proteins: Oligomerization of bacterial enterotoxins *in vivo* and *in vitro*. *Proceedings of the National Academy of Sciences of the United States of America* 1988;85:7109–7113.

74. **Sandkvist M, Michel LO, Hough LP, Morales VM, Bagdasarian M, *et al.*** General secretion pathway (*eps*) genes required for toxin secretion and outer membrane biogenesis in *Vibrio cholerae*. *Journal of Bacteriology* 1997;179:6994.

75. **Sandkvist M, Morales V, Bagdasarian M**. A protein required for secretion of cholera toxin through the outer membrane of *Vibrio cholerae*. *Gene* 1993;123:81–86.

76. **Sikora AE, Zielke RA, Lawrence DA, Andrews PC, Sandkvist M**. Proteomic analysis of the *Vibrio cholerae* type II secretome reveals new proteins, including three related serine proteases. *Journal of Biological Chemistry* 2011;286:16555–16566.

77. **Naha A, Mandal RS, Samanta P, Saha RN, Shaw S, *et al.*** Deciphering the possible role of *ctxB7* allele on higher production of cholera toxin by Haitian variant *Vibrio cholerae* O1. *PLoS Neglected Tropical Diseases* 2020;14:e0008128.

78. **Holmgren J, Lönnroth I, Månsson J, Svennerholm L**. Interaction of cholera toxin and membrane G_{M1} ganglioside of small intestine. *Proceedings of the National Academy of Sciences of the United States of America* 1975;72:2520–2524.

79. **Merritt EA, Sarfaty S, van den Akker F, L'Hoir C, Martial JA, *et al.*** Crystal structure of cholera toxin B-pentamer bound to receptor G_{M1} pentasaccharide. *Protein Science* 1994;3:166–175.

80. **Heim JB, Hodnik V, Heggelund JE, Anderluh G, Krengel U**. Crystal structures of cholera toxin in complex with fucosylated receptors point to importance of secondary binding site. *Scientific Reports* 2019;9:12243.

81. **Heggelund JE, Burschowsky D, Bjørnestad VA, Hodnik V, Anderluh G, *et al.*** High-resolution crystal structures elucidate the molecular basis of cholera blood group dependence. *PLoS Pathogens* 2016;12:e1005567.

82. **Wands AM, Fujita A, McCombs JE, Cervin J, Dedic B, *et al.*** Fucosylation and protein glycosylation create functional receptors for cholera toxin. *eLife* 2015;4:e09545.

83. **Chinnapen DJ-F, Chinnapen H, Saslowsky D, Lencer WI**. Rafting with cholera toxin: endocytosis and trafficking from plasma membrane to ER. *FEMS Microbiology Letters* 2007;266:129–137.

84. **Wernick NLB, Chinnapen DJ-F, Cho JA, Lencer WI**. Cholera toxin: An intracellular journey into the cytosol by way of the endoplasmic reticulum. *Toxins (Basel)* 2010;2:310–325.

85. **Fujinaga Y, Wolf AA, Rodighiero C, Wheeler H, Tsai B, *et al.*** Gangliosides that associate with lipid rafts mediate transport of cholera and related toxins from the plasma membrane to endoplasmic reticulm. *Molecular Biology of the Cell* 2003;14:4783–4793.

86. **Tsai B, Rapoport TA**. Unfolded cholera toxin is transferred to the ER membrane and released from protein disulfide isomerase upon oxidation by Ero1. *Journal of Cell Biology* 2002;159:207–216.

87. **Cassel D, Pfeuffer T**. Mechanism of cholera toxin action: Covalent modification of the guanyl nucleotide-binding protein of the adenylate cyclase system. *Proceedings of the National Academy of Sciences of the United States of America* 1978;75:2669–2673.

88. **Gill DM, Meren R**. ADP-ribosylation of membrane proteins catalyzed by cholera toxin: basis of the activation of adenylate cyclase. *Proceedings of the National Academy of Sciences of the United States of America* 1978;75:3050–3054.

89. **Holmgren J**. Actions of cholera toxin and the prevention and treatment of cholera. *Nature* 1981;292:413–417.

90. **Gabriel SE, Brigman KN, Koller BH, Boucher RC, Stutts MJ**. Cystic fibrosis heterozygote resistance to cholera toxin in the cystic fibrosis mouse model. *Science* 1994;266:107–109.

91. **Taylor RK, Miller VL, Furlong DB, Mekalanos JJ**. Use of *phoA* gene fusions to identify a pilus colonization factor coordinately regulated with cholera toxin. *Proceedings of the National Academy of Sciences of the United States of America* 1987;84:2833–2837.

92. **Herrington DA, Hall RH, Losonsky G, Mekalanos JJ, Taylor RK, *et al.*** Toxin, toxin-coregulated pili, and the *toxR* regulon are essential for *Vibrio cholerae* pathogenesis in humans. *Journal of Experimental Medicine* 1988;168:1487–1492.

93. **Boyd EF, Waldor MK**. Alternative mechanism of cholera toxin acquisition by *Vibrio cholerae*: Generalized transduction of CTXΦ by bacteriophage CP-T1. *Infection and Immunity* 1999;67:5898–5905.

94. **Campos J, Martínez E, Izquierdo Y, Fando R**. VEJφ, a novel filamentous phage of *Vibrio cholerae* able to transduce the cholera toxin genes. *Microbiology* 2010;156:108–115.

95. **McLeod SM, Waldor MK**. Characterization of XerC- and XerD-dependent CTX phage integration in *Vibrio cholerae*. *Molecular Microbiology* 2004;54:935–947.

96. **Huber KE, Waldor MK**. Filamentous phage integration requires the host recombinases XerC and XerD. *Nature* 2002;417:656.

97. **Feng L, Reeves PR, Lan R, Ren Y, Gao C, *et al.*** A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS ONE* 2008;3:e4053.

98. **Grim CJ, Hasan NA, Taviani E, Haley B, Chun J, *et al.*** Genome sequence of hybrid *Vibrio cholerae* O1 MJ-1236, B-33, and CIRS101 and comparative genomics with *V. cholerae*. *Journal of Bacteriology* 2010;192:3524–3533.

99. **McLeod SM, Kimsey HH, Davis BM, Waldor MK**. CTXφ and *Vibrio cholerae*: Exploring a newly recognized type of phage–host cell relationship. *Molecular Microbiology* 2005;57:347–356.

100. **Moyer KE, Kimsey HH, Waldor MK**. Evidence for a rolling-circle mechanism of phage DNA synthesis from both replicative and integrated forms of CTXφ. *Molecular Microbiology* 2001;41:311–323.

101. **Davis BM, Waldor MK**. CTXφ contains a hybrid genome derived from tandemly integrated elements. *Proceedings of the National Academy of Sciences of the United States of America* 2000;97:8572–8577.

102. **Miller V, Mekalanos J**. Synthesis of cholera toxin is positively regulated at the transcriptional level by *toxR*. *Proceedings of the National Academy of Sciences of the United States of America* 1984;81:3471–3475.

103. **Holmgren J, Svennerholm A-M**. Mechanisms of disease and immunity in cholera: A review. *Journal of Infectious Diseases* 1977;136:S105–S112.

104. **Almagro-Moreno S, Pruss K, Taylor RK**. Intestinal colonization dynamics of *Vibrio cholerae*. *PLoS Pathogens* 2015;11:e1004787.

105. **Miller VL, Taylor RK, Mekalanos JJ**. Cholera toxin transcriptional activator ToxR is a transmembrane DNA binding protein. *Cell* 1987;48:271–279.

106. **Higgins DE, Nazareno E, DiRita VJ**. The virulence gene activator ToxT from *Vibrio cholerae* is a member of the AraC family of transcriptional activators. *Journal of Bacteriology* 1992;174:6974–6980.

107. **Withey JH, DiRita VJ**. The toxbox: specific DNA sequence requirements for activation of *Vibrio cholerae* virulence genes by ToxT. *Molecular Microbiology* 2006;59:1779–1789.

108. **Miller VL, Mekalanos JJ**. Genetic analysis of the cholera toxin-positive regulatory gene *toxR*. *Journal of Bacteriology* 1985;163:580–585.

109. **DiRita VJ, Parsot C, Jander G, Mekalanos JJ**. Regulatory cascade controls virulence in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 1991;88:5403–5407.

110. **Ottemann KM, Mekalanos JJ**. The ToxR protein of *Vibrio cholerae* forms homodimers and heterodimers. *Journal of Bacteriology* 1996;178:156–162.

111. **Withey JH, DiRita VJ**. Activation of both *acfA* and *acfD* transcription by *Vibrio cholerae* ToxT requires binding to two centrally located DNA sites in an inverted repeat conformation. *Molecular Microbiology* 2005;56:1062–1077.

112. **Higgins DE, DiRita VJ**. Transcriptional control of *toxT*, a regulatory gene in the ToxR regulon of *Vibrio cholerae*. *Molecular Microbiology* 1994;14:17–29.

113. **Häse CC, Mekalanos JJ**. TcpP protein is a positive regulator of virulence gene expression in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 1998;95:730–734.

114. **Carroll PA, Tashima KT, Rogers MB, DiRita VJ, Calderwood SB**. Phase variation in *tcpH* modulates expression of the ToxR regulon in *Vibrio cholerae*. *Molecular Microbiology* 1997;25:1099–1111.

115. **Beck NA, Krukonis ES, DiRita VJ**. TcpH influences virulence gene expression in *Vibrio cholerae* by inhibiting degradation of the transcription activator TcpP. *Journal of Bacteriology* 2004;186:8309–8316.

116. **Skorupski K, Taylor RK**. A new level in the *Vibrio cholerae* ToxR virulence cascade: AphA is required for transcriptional activation of the *tcpPH* operon. *Molecular Microbiology* 1999;31:763–771.

117. **Behari J, Stagon L, Calderwood SB**. *pepA*, a gene mediating pH regulation of virulence genes in *Vibrio cholerae*. *Journal of Bacteriology* 2001;183:178–188.

118. **Gupta S, Chowdhury R**. Bile affects production of virulence factors and motility of *Vibrio cholerae*. *Infection and Immunity* 1997;65:1131–1134.

119. **Yang M, Liu Z, Hughes C, Stern AM, Wang H, *et al.*** Bile salt–induced intermolecular disulfide bond formation activates *Vibrio cholerae* virulence. *Proceedings of the National Academy of Sciences of the United States of America* 2013;110:2348–2353.

120. **Bina XR, Taylor DL, Vikram A, Ante VM, Bina JE**. *Vibrio cholerae* ToxR downregulates virulence factor production in response to cyclo(Phe-Pro). *mBio* 2013;4:e00366-13.

121. **D'Haeze W**. Three-way regulation of cholera toxin production. *Genome Biology* 2002;3:reports0056.

122. **Tischler AD, Lee SH, Camilli A**. The *Vibrio cholerae vieSAB* locus encodes a pathway contributing to cholera toxin production. *Journal of Bacteriology* 2002;184:4104–4113.

123. **Ayala JC, Wang H, Benitez JA, Silva AJ**. Molecular basis for the differential expression of the global regulator VieA in *Vibrio cholerae* biotypes directed by H-NS, LeuO and quorum sensing. *Molecular Microbiology* 2018;107:330–343.

124. **Conner JG, Zamorano-Sánchez D, Park JH, Sondermann H, Yildiz FH**. The ins and outs of cyclic di-GMP signaling in *Vibrio cholerae*. *Current Opinion in Microbiology* 2017;36:20–29.

125. **Tamayo R, Schild S, Pratt JT, Camilli A**. Role of cyclic di-GMP during El Tor biotype *Vibrio cholerae* infection: Characterization of the *in vivo*-induced cyclic di-GMP phosphodiesterase CdpA. *Infection and Immunity* 2008;76:1617–1627.

126. **Sakib SN, Reddi G, Almagro-Moreno S**. Environmental role of pathogenic traits in *Vibrio cholerae*. *Journal of Bacteriology* 2018;200:e00795-17.

127. **Schild S, Tamayo R, Nelson EJ, Qadri F, Calderwood SB, *et al.*** Genes induced late in infection increase fitness of *Vibrio cholerae* after release into the environment. *Cell Host and Microbe* 2007;2:264–277.

128. **LaRocque RC, Harris JB, Dziejman M, Li X, Khan AI, *et al.*** Transcriptional profiling of *Vibrio cholerae* recovered directly from patient specimens during early and late stages of human infection. *Infection and Immunity* 2005;73:4488–4493.

129. **Bina J, Zhu J, Dziejman M, Faruque S, Calderwood S, *et al.*** ToxR regulon of *Vibrio cholerae* and its expression in vibrios shed by cholera patients. *Proceedings of the National Academy of Sciences of the United States of America* 2003;100:2801–2806.

130. **Kovacikova G, Skorupski K**. The alternative sigma factor σ^E plays an important role in intestinal survival and virulence in *Vibrio cholerae*. *Infection and Immunity* 2002;70:5355–5362.

131. **Merrell DS, Tischler AD, Lee SH, Camilli A**. *Vibrio cholerae* requires *rpoS* for efficient intestinal colonization. *Infection and Immunity* 2000;68:6691–6696.

132. **Silva AJ, Benitez JA**. *Vibrio cholerae* biofilms and cholera pathogenesis. *PLoS Neglected Tropical Diseases* 2016;10:e0004330.

133. **Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, *et al.*** Comparative genomic analysis of *Vibrio cholerae*: Genes that correlate with cholera endemic and pandemic disease. *Proceedings of the National Academy of Sciences of the United States of America* 2002;99:1556–1561.

134. **Jermyn WS, Boyd EF**. Characterization of a novel *Vibrio* pathogenicity island (VPI-2) encoding neuraminidase (*nanH*) among toxigenic *Vibrio cholerae* isolates. *Microbiology* 2002;148:3681–3693.

135. **Burnet FM, Stone JD**. The receptor-destroying enzyme of *V. cholerae*. *Australian Journal of Experimental Biology and Medical Science* 1947;25:227–233.

136. **Vimr ER, Lawrisuk L, Galen J, Kaper JB**. Cloning and expression of the *Vibrio cholerae* neuraminidase gene *nanH* in *Escherichia coli*. *Journal of Bacteriology* 1988;170:1495–1504.

137. **Svennerholm L**. Chromatographic separation of human brain gangliosides. *Journal of Neurochemistry* 1963;10:613–623.

138. **Almagro-Moreno S, Boyd EF**. Sialic acid catabolism confers a competitive advantage to pathogenic *Vibrio cholerae* in the mouse intestine. *Infection and Immunity* 2009;77:3807–3816.

139. **Labbate M, Orata FD, Petty NK, Jayatilleke ND, King WL, *et al.*** A genomic island in *Vibrio cholerae* with VPI-1 site-specific recombination characteristics contains CRISPR-Cas and type VI secretion modules. *Scientific Reports* 2016;6:36891.

140. **Rapa RA, Islam A, Monahan LG, Mutreja A, Thomson N, *et al.*** A genomic island integrated into *recA* of *Vibrio cholerae* contains a divergent *recA* and provides multi-pathway protection from DNA damage. *Environmental Microbiology* 2015;17:1090–1102.

141. **Murphy RA, Boyd EF**. Three pathogenicity islands of *Vibrio cholerae* can excise from the chromosome and form circular intermediates. *Journal of Bacteriology* 2008;190:636–647.

142. **Davies BW, Bogard RW, Young TS, Mekalanos JJ**. Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for *V. cholerae* virulence. *Cell* 2012;149:358–370.

143. **Imamura D, Morita M, Sekizuka T, Mizuno T, Takemura T, *et al.*** Comparative genome analysis of VSP-II and SNPs reveals heterogenic variation in contemporary strains of *Vibrio cholerae* O1 isolated from cholera patients in Kolkata, India. *PLoS Neglected Tropical Diseases* 2017;11:e0005386.

144. **Blokesch M, Schoolnik GK**. Serogroup conversion of *Vibrio cholerae* in aquatic reservoirs. *PLoS Pathogens* 2007;3:e81.

145. **Bik EM, Bunschoten AE, Gouw RD, Mooi FR**. Genesis of the novel epidemic *Vibrio cholerae* O139 strain: Evidence for horizontal transfer of genes involved in polysaccharide synthesis. *The EMBO Journal* 1995;14:209–216.

146. **Albert MJ, Alam K, Ansaruzzaman M, Qadri F, Sack RB**. Lack of cross-protection against diarrhea due to *Vibrio cholerae* O139 (Bengal strain) after oral immunization of rabbits with *V. cholerae* O1 vaccine strain CVD103-HgR. *Journal of Infectious Diseases* 1994;169:230–231.

147. **Davies J, Davies D**. Origins and evolution of antibiotic resistance. *Microbiology and Molecular Biology Reviews* 2010;74:417–433.

148. **Carraro N, Rivard N, Ceccarelli D, Colwell RR, Burrus V**. IncA/C conjugative plasmids mobilize a new family of multidrug resistance islands in clinical *Vibrio cholerae* non-O1/non-O139 isolates from Haiti. *mBio* 2016;7:e00509-16.

149. **Folster JP, Katz L, McCullough A, Parsons MB, Knipe K, *et al.*** Multidrug-resistant IncA/C plasmid in *Vibrio cholerae* from Haiti. *Emerging Infectious Diseases* 2014;20:1951–1952.

150. **Fonseca ÉL, dos Santos Freitas F, Vieira VV, Vicente ACP**. New *qnr* gene cassettes associated with superintegron repeats in *Vibrio cholerae* O1. *Emerging Infectious Diseases* 2008;14:1129–1131.

151. **Petroni A, Melano RG, Saka HA, Garutti A, Mange L, *et al.*** CARB-9, a carbenicillinase encoded in the VCR region of *Vibrio cholerae* non-O1, non-O139 belongs to a family of cassette-encoded β-lactamases. *Antimicrobial Agents and Chemotherapy* 2004;48:4042–4046.

152. **Melano R, Petroni A, Garutti A, Saka HA, Mange L, *et al.*** New carbenicillin-hydrolyzing β-lactamase (CARB-7) from *Vibrio cholerae* non-O1, non-O139 strains encoded by the VCR region of the *V. cholerae* genome. *Antimicrobial Agents and Chemotherapy* 2002;46:2162–2168.

153. **Kim HB, Wang M, Ahmed S, Park CH, LaRocque RC, *et al.*** Transferable quinolone resistance in *Vibrio cholerae*. *Antimicrobial Agents and Chemotherapy* 2010;54:799–803.

154. **Fonseca ÉL da, Vicente ACP**. Spread of the *qnrVC* quinolone resistance determinant in *Vibrio cholerae*. *Antimicrobial Agents and Chemotherapy* 2011;55:457–457.

155. **Zhou Y, Yu L, Li J, Zhang L, Tong Y, *et al.*** Accumulation of mutations in DNA gyrase and topoisomerase IV genes contributes to fluoroquinolone resistance in *Vibrio cholerae* O139 strains. *International Journal of Antimicrobial Agents* 2013;42:72–75.

156. **Waldor MK, Tschäpe H, Mekalanos JJ**. A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* O139. *Journal of Bacteriology* 1996;178:4157–4165.

157. **Towner KJ, Pearson NJ, Mhalu FS, O'Grady F**. Resistance to antimicrobial agents of *Vibrio cholerae* El Tor strains isolated during the fourth cholera epidemic in the United Republic of Tanzania. *Bulletin of the World Health Organization* 1980;58:747–751.

158. **Weill F-X, Domman D, Njamkepo E, Tarr C, Rauzier J, *et al.*** Genomic history of the seventh pandemic of cholera in Africa. *Science* 2017;358:785–789.

159. **Kelly H**. The classical definition of a pandemic is not elusive. *Bulletin of the World Health Organization* 2011;89:540–541.

160. **Last J**. *A dictionary of epidemiology*. 4th ed. New York, NY: Oxford University Press; 2001.

161. **Barua D**. The global epidemiology of cholera in recent years. *Proceedings of the Royal Society of Medicine* 1972;65:423–428.

162. **Devault AM, Golding GB, Waglechner N, Enk JM, Kuch M, *et al.*** Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. *New England Journal of Medicine* 2014;370:334–340.

163. **Salim A, Lan R, Reeves PR**. *Vibrio cholerae* pathogenic clones. *Emerging Infectious Diseases* 2005;11:1758–1760.

164. **Cvjetanovic B, Barua D**. The seventh pandemic of cholera. *Nature* 1972;239:137–138.

165. **World Health Organization Regional Office for the Western Pacific**. *Meeting for the Exchange of Information on El Tor Vibrio Paracholera, Manila, Philippines, 16-19 April 1962 : final report*. Technical Report; Manila: WHO Regional Office for the Western Pacific. https://iris.wpro.who.int/handle/10665.1/5806 (1962, accessed 30 June 2020).

166. **De Moor EC**. Paracholera (El Tor): Enteritis choleriformis El Tor van Loghem. *Bulletin of the World Health Organization* 1949;2:5–17.

167. **Rebaudet S, Sudre B, Faucher B, Piarroux R**. Environmental determinants of cholera outbreaks in inland Africa: A systematic review of main transmission foci and propagation routes. *Journal of Infectious Diseases* 2013;208:S46–S54.

168. **Kumate J, Sepúlveda J, Gutiérrez G**. Cholera epidemiology in Latin America and perspectives for eradication. *Bulletin de l'Institut Pasteur* 1998;96:217–226.

169. **Barua D**. History of cholera. In: Barua D, Greenough WB (editors). *Cholera*. Boston, MA: Springer US. pp. 1–36.

170. **van de Linde PAM, Forbes GI**. Observations on the spread of cholera in Hong Kong, 1961-63. *Bulletin of the World Health Organization* 1965;32:515–530.

171. **Rebaudet S, Sudre B, Faucher B, Piarroux R**. Cholera in coastal Africa: A systematic review of its heterogeneous environmental determinants. *Journal of Infectious Diseases* 2013;208:S98–S106.

172. **Swerdlow DL, Isaäcson M**. The epidemiology of cholera in Africa. *Vibrio cholerae and Cholera* 1994;297–307.

173. **Centers for Disease Control and Prevention (CDC)**. Cholera--Peru, 1991. *Morbidity and Mortality Weekly Report* 1991;40:108–110.

174. **Tauxe RV, Mintz ED, Quick RE**. Epidemic cholera in the new world: Translating field epidemiology into new prevention strategies. *Emerging Infectious Diseases* 1995;1:141–146.

175. **Guthmann JP**. Epidemic cholera in Latin America: Spread and routes of transmission. *Journal of Tropical Medicine and Hygiene* 1995;98:419–427.

176. **Centers for Disease Control and Prevention (CDC)**. Cholera outbreak --- Haiti, October 2010. *Morbidity and Mortality Weekly Report* 2010;59:1411.

177. **Centers for Disease Control and Prevention (CDC)**. Update on cholera --- Haiti, Dominican Republic, and Florida, 2010. *Morbidity and Mortality Weekly Report* 2010;59:1637–1641.

178. **Dowell SF, Tappero JW, Frieden TR**. Public health in Haiti — Challenges and progress. *New England Journal of Medicine* 2011;364:300–301.

179. **Guillaume Y, Ternier R, Vissieres K, Casseus A, Chery MJ, *et al.*** Responding to cholera in Haiti: Implications for the national plan to eliminate cholera by 2022. *Journal of Infectious Diseases* 2018;218:S167–S170.

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

181. **Huq A, Colwell RR, Rahman R, Ali A, Chowdhury MA, *et al.*** Detection of *Vibrio cholerae* O1 in the aquatic environment by fluorescent-monoclonal antibody and culture methods. *Applied and Environmental Microbiology* 1990;56:2370–2373.

182. **Colwell R, Huq A**. Marine ecosystems and cholera. *Hydrobiologia* 2001;460:141–145.

183. **Islam MS, Drasar BS, Bradley DJ**. Long-term persistence of toxigenic *Vibrio cholerae* 01 in the mucilaginous sheath of a blue-green alga, *Anabaena variabilis*. *Journal of Tropical Medicine and Hygiene* 1990;93:133–139.

184. **Alam M, Hasan NA, Sadique A, Bhuiyan NA, Ahmed KU, *et al.*** Seasonal cholera caused by *Vibrio cholerae* serogroups O1 and O139 in the coastal aquatic environment of Bangladesh. *Applied and Environmental Microbiology* 2006;72:4096–4104.

185. **Xu H-S, Roberts N, Singleton FL, Attwell RW, Grimes DJ, *et al.*** Survival and viability of nonculturable *Escherichia coli* and *Vibrio cholerae* in the estuarine and marine environment. *Microbial Ecology* 1982;8:313–323.

186. **Binsztein N, Costagliola MC, Pichel M, Jurquiza V, Ramírez FC, *et al.*** Viable but nonculturable *Vibrio cholerae* O1 in the aquatic environment of Argentina. *Applied and Environmental Microbiology* 2004;70:7481–7486.

187. **Kaper J, Lockman H, Colwell RR, Joseph SW**. Ecology, serology, and enterotoxin production of *Vibrio cholerae* in Chesapeake Bay. *Applied and Environmental Microbiology* 1979;37:91–103.

188. **Vezzulli L, Pruzzo C, Huq A, Colwell RR**. Environmental reservoirs of *Vibrio cholerae* and their role in cholera. *Environmental Microbiology Reports* 2010;2:27–33.

189. **Domman D, Quilici ML, Dorman MJ, Njamkepo E, Mutreja A, *et al.*** Integrated view of *Vibrio cholerae* in the Americas. *Science* 2017;358:789–793.

190. **Shimada T, Sakazaki R**. Additional serovars and inter-O antigenic relationships of *Vibrio cholerae*. *Japanese Journal of Medical Science and Biology* 1977;30:275–277.

191. **Svennerholm AM, Holmgren J**. Synergistic protective effect in rabbits of immunization with *Vibrio cholerae* lipopolysaccharide and toxin/toxoid. *Infection and Immunity* 1976;13:735–740.

192. **Gardner AD, Venkatraman KV**. The antigens of *Vibrio cholerae*. *The Lancet* 1935;225:265.

193. **Aoki Y**. A request for unification in type and antigen designations of *Vibrio cholerae*. *Japanese Journal of Microbiology* 1962;6:79–82.

194. **Gustafsson B**. Monoclonal antibody-based enzyme-linked immunosorbent assays for identification and serotyping of *Vibrio cholerae* O1. *Journal of Clinical Microbiology* 1984;20:1180–1185.

195. **Gustafsson B, Holme T**. Immunological characterization of *Vibrio cholerae* O:1 lipopolysaccharide, O-side chain, and core with monoclonal antibodies. *Infection and Immunity* 1985;49:275–280.

196. **Lebens M, Karlsson SL, Källgård S, Blomquist M, Ekman A, *et al.*** Construction of novel vaccine strains of *Vibrio cholerae* co-expressing the Inaba and Ogawa serotype antigens. *Vaccine* 2011;29:7505–7513.

197. **Stroeher UH, Karageorgos LE, Morona R, Manning PA**. Serotype conversion in *Vibrio cholerae* O1. *Proceedings of the National Academy of Sciences of the United States of America* 1992;89:2566–2570.

198. **Ito T, Hiramatsu K, Ohshita Y, Yokota T**. Mutations in the *rfbT* gene are responsible for the Ogawa to Inaba serotype conversion in *Vibrio cholerae* O1. *Microbiology and Immunology* 1993;37:281–288.

199. **Chatterjee SN, Chaudhuri K**. Lipopolysaccharides of *Vibrio cholerae*. I. Physical and chemical characterization. *Biochimica et Biophysica Acta* 2003;1639:65–79.

200. **Karlsson SL, Thomson N, Mutreja A, Connor T, Sur D, *et al.*** Retrospective analysis of serotype switching of *Vibrio cholerae* O1 in a cholera endemic region shows it is a non-random process. *PLoS Neglected Tropical Diseases* 2016;10:e0005044.

201. **Sheehy TW, Sprinz H, Augerson WS, Formal SB**. Laboratory *Vibrio cholerae* infection in the United States. *Journal of the American Medical Association* 1966;197:321–326.

202. **Sack RB, Miller CE**. Progressive changes of *Vibrio* serotypes in germ-free mice infected with *Vibrio cholerae*. *Journal of Bacteriology* 1969;99:688–695.

203. **Khan AI, Chowdhury F, Harris JB, Larocque RC, Faruque ASG, *et al.*** Comparison of clinical features and immunological parameters of patients with dehydrating diarrhoea infected with Inaba or Ogawa serotypes of *Vibrio cholerae* O1. *Scandinavian Journal of Infectious Disease* 2010;42:48–56.

204. **European Medicines Agency**. Dukoral. *European Medicines Agency*. https://www.ema.europa.eu/en/medicines/human/EPAR/dukoral (2018, accessed 13 November 2019).

205. **Karlsson SL, Ax E, Nygren E, Källgård S, Blomquist M, *et al.*** Development of stable *Vibrio cholerae* O1 Hikojima type vaccine strains co–expressing the Inaba and Ogawa lipopolysaccharide antigens. *PLoS ONE* 2014;9:e108521.

206. **Centers for Disease Control and Prevention (CDC)**. Chapter 6 – Laboratory identification of *Vibrio cholerae*. In: *Laboratory methods for the diagnosis of* Vibrio cholerae. Centers for Disease Control and Prevention.

207. **Alam MT, Ray SS, Chun CN, Chowdhury ZG, Rashid MH, *et al.*** Major shift of toxigenic *V. cholerae* O1 from Ogawa to Inaba serotype isolated from clinical and environmental samples in Haiti. *PLoS Neglected Tropical Diseases* 2016;10:e0005045.

208. **Vugia DJ, Rodriguez M, Vargas R, Ricse C, Ocampo C, *et al.*** Epidemic cholera in Trujillo, Peru 1992: Utility of a clinical case definition and shift in *Vibrio cholerae* O1 serotype. *The American Journal of Tropical Medicine and Hygiene* 1994;50:566–569.

209. **Hugh R**. The proposed conservation of the generic name *Vibrio* Pacini 1854 and designation of the neotype strain of *Vibrio cholerae* Pacini 1854. *International Journal of Systematic and Evolutionary Microbiology* 1964;14:87–101.

210. **Hugh R**. A comparison of *Vibrio cholerae* Pacini and *Vibrio eltor* Pribram. *International Journal of Systematic and Evolutionary Microbiology,* 1965;15:61–68.

211. **Gotschlich F**. Über cholera- und choleraähnliche Vibrionen unter den aus Mekka zurückkehrenden Pilgern. *Zeitschrift für Hygiene und Infektionskrankheiten* 1906;53:281–304.

212. **Ruffer MA**. Researches on the bacteriological diagnosis of cholera, carried out by medical officers of the sanitary, maritime and quarantine council of Egypt. *The British Medical Journal* 1907;1:735–742.

213. **Hu D, Liu B, Feng L, Ding P, Guo X, *et al.*** Origins of the current seventh cholera pandemic. *Proceedings of the National Academy of Sciences of the United States of America* 2016;113:E7730–E7739.

214. **Castellani A**. Paracholera. *British Medical Journal* 1916;1:448–449.

215. **Mackie TJ, Storer EJ**. Two *Vibrio* species of the "paracholera" group associated with a cholera-like outbreak. *Journal of the Royal Army Medical Corps* 1918;31:161–169.

216. **Finkelstein MH**. Problems in the bacteriology of cholera and cholera-like infections. *Transactions of the Royal Society of Tropical Medicine and Hygiene* 1931;25:29–38.

217. **Mackie TJ**. The serological relationships of the paracholera Vibrios to *Vibrio choleræ*, and the serological races of the paracholera group. *British Journal of Experimental Pathology* 1922;3:231–237.

218. **van Loghem JJ**. Über den unterschied zwischen cholera- und El Tor-Vibrionen. *Zentralblatt für Bakteriologie*;67.

219. **Alm RA, Stroeher UH, Manning PA**. Extracellular proteins of *Vibrio cholerae*: nucleotide sequence of the structural gene (*hlyA*) for the haemolysin of the haemolytic El Tor strain 017 and characterization of the *hlyA* mutation in the non-haemolytic classical strain 569B. *Molecular Microbiology* 1988;2:481–488.

220. **Manning PA, Brown MH, Heuzenroeder MW**. Cloning of the structural gene (*hly*) for the haemolysin of *Vibrio cholerae* El Tor strain 017. *Gene* 1984;31:225–231.

221. **Fan Y, Li Z, Li Z, Li X, Sun H, *et al.*** Nonhemolysis of epidemic El Tor biotype strains of *Vibrio cholerae* is related to multiple functional deficiencies of hemolysin A. *Gut Pathogens* 2019;11:38.

222. **Son MS, Megli CJ, Kovacikova G, Qadri F, Taylor RK**. Characterization of *Vibrio cholerae* O1 El Tor biotype variant clinical isolates from Bangladesh and Haiti, including a molecular genetic analysis of virulence genes. *Journal of Clinical Microbiology* 2011;49:3739–3749.

223. **Albert MJ**. *Vibrio cholerae* O139 Bengal. *Journal of Clinical Microbiology* 1994;32:2345–2349.

224. **Cholera Working Group for ICDDR Bangladesh, Albert MJ, Ansaruzzaman M, Bardhan PK, Faruque ASG, *et al.*** Large epidemic of cholera-like disease in Bangladesh caused by *Vibrio cholerae* 0139 synonym Bengal. *The Lancet* 1993;342:387–390.

225. **Ramamurthy T, Garg S, Sharma R, Bhattacharya SK, Nair GB, *et al.*** Emergence of novel strain of *Vibrio cholerae* with epidemic potential in southern and eastern India. *The Lancet* 1993;341:703–704.

226. **Finkelstein RA**. Cholera, *Vibrio cholerae* O1 and O139, and other pathogenic Vibrios. In: Baron S (editor). *Medical Microbiology*. Galveston (TX): University of Texas Medical Branch at Galveston. http://www.ncbi.nlm.nih.gov/books/NBK8407/ (1996).

227. **World Health Organization**. WHO | 1998 - Cholera - *Vibrio cholerae* O139 strain. *World Health Organization*. http://www.who.int/csr/don/1998_09_22/en/ (accessed 3 January 2018).

228. **Faruque AS, Fuchs GJ, Albert MJ**. Changing epidemiology of cholera due to *Vibrio cholerae* O1 and O139 Bengal in Dhaka, Bangladesh. *Epidemiology and Infection* 1996;116:275–278.

229. **Faruque SM, Chowdhury N, Kamruzzaman M, Ahmad QS, Faruque ASG, *et al.*** Reemergence of epidemic *Vibrio cholerae* O139, Bangladesh. *Emerging Infectious Diseases* 2003;9:1116–1122.

230. **Faruque SM, Sack DA, Sack RB, Colwell RR, Takeda Y, *et al.*** Emergence and evolution of *Vibrio cholerae* O139. *Proceedings of the National Academy of Sciences of the United States of America* 2003;100:1304–1309.

231. **Berche P, Poyart C, Abachin E, Lelievre H, Vandepitte J, *et al.*** The novel epidemic strain O139 is closely related to the pandemic strain O1 of *Vibrio cholerae*. *Journal of Infectious Diseases* 1994;170:701–704.

232. **Calia KE, Waldor MK, Calderwood SB**. Use of representational difference analysis to identify genomic differences between pathogenic strains of *Vibrio cholerae*. *Infection and Immunity* 1998;66:849–852.

233. **Hall RH, Khambaty FM, Kothary MH, Keasler SP, Tall BD**. *Vibrio cholerae* non-O1 serogroup associated with cholera gravis genetically and physiologically resembles O1 E1 Tor cholera strains. *Infection and Immunity* 1994;62:3859–3863.

234. **Mutreja A, Kim DW, Thomson NR, Connor TR, Lee JH, *et al.*** Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* 2011;477:462–465.

235. **Domman D, Chowdhury F, Khan AI, Dorman MJ, Mutreja A, *et al.*** Defining endemic cholera at three levels of spatiotemporal resolution within Bangladesh. *Nature Genetics* 2018;50:951–955.

236. **Weintraub A, Widmalm G, Jansson P-E, Jansson M, Hultenby K, *et al.*** *Vibrio cholerae* O139 Bengal possesses a capsular polysaccharide which may confer increased virulence. *Microbial Pathogenesis* 1994;16:235–241.

237. **Sozhamannan S, Deng YK, Li M, Sulakvelidze A, Kaper JB, *et al.*** Cloning and sequencing of the genes downstream of the *wbf* gene cluster of *Vibrio cholerae* serogroup O139 and analysis of the junction genes in other serogroups. *Infection and Immunity* 1999;67:5033–5040.

238. **Stroeher UH, Parasivam G, Dredge BK, Manning PA**. Novel *Vibrio cholerae* O139 genes involved in lipopolysaccharide biosynthesis. *Journal of Bacteriology* 1997;179:2740–2747.

239. **Yamasaki S, Garg S, Nair GB, Takeda Y**. Distribution of *Vibrio cholerae* O1 antigen biosynthesis genes among O139 and other non-O1 serogroups of *Vibrio cholerae*. *FEMS Microbiology Letters* 1999;179:115–121.

240. **Waldor MK, Colwell R, Mekalanos JJ**. The *Vibrio cholerae* O139 serogroup antigen includes an O-antigen capsule and lipopolysaccharide virulence determinants. *Proceedings of the National Academy of Sciences of the United States of America* 1994;91:11388–11392.

241. **Waldor MK, Mekalanos JJ**. ToxR regulates virulence gene expression in non-O1 strains of *Vibrio cholerae* that cause epidemic cholera. *Infection and Immunity* 1994;62:72–78.

242. **Siriphap A, Leekitcharoenphon P, Kaas RS, Theethakaew C, Aarestrup FM, *et al.*** Characterization and genetic variation of *Vibrio cholerae* isolated from clinical and environmental sources in Thailand. *PLoS ONE* 2017;12:e0169324.

243. **Yi Y, Lu N, Liu F, Li J, Zhang R, *et al.*** Genome sequence and comparative analysis of a *Vibrio cholerae* O139 strain E306 isolated from a cholera case in China. *Gut Pathogens* 2014;6:3.

244. **Chowdhury F, Mather AE, Begum YA, Asaduzzaman M, Baby N, *et al.*** *Vibrio cholerae* serogroup O139: Isolation from cholera patients and asymptomatic household family members in Bangladesh between 2013 and 2014. *PLoS Neglected Tropical Diseases* 2015;9:e0004183.

245. **World Health Organization**. Weekly epidemiological record. *Weekly Epidemiological Record* 2004;79:281-288.

246. **World Health Organization**. Weekly epidemiological record. *Weekly Epidemiological Record* 2017;92:477–500.

247. **Saha A, Chowdhury MI, Khanam F, Bhuiyan MdS, Chowdhury F, *et al.*** Safety and immunogenicity study of a killed bivalent (O1 and O139) whole-cell oral cholera vaccine Shanchol, in Bangladeshi adults and children as young as 1 year of age. *Vaccine* 2011;29:8285–8292.

248. **World Health Organization**. Weekly epidemiological record. *Weekly Epidemiological Record* 1969;44:1–27.

249. **Heiberg B**. The biochemical reactions of Vibrios. *Journal of Hygiene (London)* 1936;36:114–117.

250. **Aldová E, Lázničková K, Ġtěpánková E, Lietaya J**. Isolation of nonagglutinable Vibrios from an enteritis outbreak in Czechoslovakia. *Journal of Infectious Diseases* 1968;118:25–31.

251. **Felsenfeld O, Stegherr-Barrios A, Aldová E, Holmes J, Parrott MW**. *In vitro* and *in vivo* studies of streptomycin-dependent cholera Vibrios. *Applied Microbiology* 1970;19:463–469.

252. **American Type Culture Collections**. *Vibrio cholerae* Pacini ATCC ® 25872™. https://www.lgcstandards-atcc.org/Products/All/25872.aspx?geo_country=gb#history (accessed 28 June 2020).

253. **Boyd EF, Heilpern AJ, Waldor MK**. Molecular analyses of a putative CTXφ precursor and evidence for independent acquisition of distinct CTXφs by toxigenic *Vibrio cholerae*. *Journal of Bacteriology* 2000;182:5530–5538.

254. **Sakazaki R, Tamura K, Gomez CZ, Sen R**. Serological studies on the cholera group of vibrios. *Japanese Journal of Medical Science and Biology* 1970;23:13–20.

255. **Chapman C, Henry M, Bishop-Lilly KA, Awosika J, Briska A, *et al.*** Scanning the landscape of genome architecture of non-O1 and non-O139 *Vibrio cholerae* by whole genome mapping reveals extensive population genetic diversity. *PLoS ONE* 2015;10:e0120311.

256. **Zinnaka Y, Carpenter CC**. An enterotoxin produced by noncholera vibrios. *Johns Hopkins Medical Journal* 1972;131:403–411.

257. **Boyd EF, Waldor MK**. Evolutionary and functional analyses of variants of the toxin-coregulated pilus protein TcpA from toxigenic *Vibrio cholerae* non-O1/non-O139 serogroup isolates. *Microbiology,* 2002;148:1655–1666.

258. **Henst CV der, Vanhove AS, Dörr NCD, Stutzmann S, Stoudmann C, *et al.*** Molecular insights into *Vibrio cholerae*'s intra-amoebal host-pathogen interactions. *Nature Communications* 2018;9:1–13.

259. **Bik EM, Gouw RD, Mooi FR**. DNA fingerprinting of *Vibrio cholerae* strains with a novel insertion sequence element: A tool to identify epidemic strains. *Journal of Clinical Microbiology* 1996;34:1453–1461.

260. **O'Shea YA, Reen FJ, Quirke AM, Boyd EF**. Evolutionary genetic analysis of the emergence of epidemic *Vibrio cholerae* isolates on the basis of comparative nucleotide sequence analysis and multilocus virulence gene profiles. *Journal of Clinical Microbiology* 2004;42:4657–4671.

261. **Beltrán P, Delgado G, Navarro A, Trujillo F, Selander RK, *et al.*** Genetic diversity and population structure of *Vibrio cholerae*. *Journal of Clinical Microbiology* 1999;37:581–590.

262. **Reen FJ, Boyd EF**. Molecular typing of epidemic and nonepidemic *Vibrio cholerae* isolates and differentiation of *V. cholerae* and *V. mimicus* isolates by PCR–single-strand conformation polymorphism analysis. *Journal of Applied Microbiology* 2005;98:544–555.

263. **Yamamoto K, Takeda Y, Miwatani T, Craig JP**. Evidence that a non-O1 *Vibrio cholerae* produces enterotoxin that is similar but not identical to cholera enterotoxin. *Infection and Immunity* 1983;41:896–901.

264. **Li M, Shimada T, Morris JG, Sulakvelidze A, Sozhamannan S**. Evidence for the emergence of non-O1 and non-O139 *Vibrio cholerae* strains with pathogenic potential by exchange of O-antigen biosynthesis regions. *Infection and Immunity* 2002;70:2441–2453.

265. **Aydanian A, Tang L, Chen Y, Morris JG, Olsen P, *et al.*** Genetic relatedness of selected clinical and environmental non-O1/O139 *Vibrio cholerae*. *International Journal of Infectious Diseases* 2015;37:152–158.

266. **Dziejman M, Serruto D, Tam VC, Sturtevant D, Diraphat P, *et al.*** Genomic characterization of non-O1, non-O139 *Vibrio cholerae* reveals genes for a type III secretion system. *Proceedings of the National Academy of Sciences of the United States of America* 2005;102:3465–3470.

267. **Chun J, Grim CJ, Hasan NA, Lee JH, Choi SY, *et al.*** Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 2009;106:15442–15447.

268. **Farina C, Marini F, Schiaffino E, Luzzi I, Dionisi AM, *et al.*** A fatal *Vibrio cholerae* O37 enteritis. *Journal of Medical Microbiology,* 2010;59:1538–1540.

269. **Kaper JB, Bradford HB, Roberts NC, Falkow S**. Molecular epidemiology of *Vibrio cholerae* in the U.S. Gulf Coast. *Journal of Clinical Microbiology* 1982;16:129–134.

270. **Blake PA, Allegra DT, Snyder JD, Barrett TJ, McFarland L, *et al.*** Cholera — a possible endemic focus in the United States. *New England Journal of Medicine* 1980;302:305–309.

271. **Johnston JM, Martin DL, Perdue J, McFarland LM, Caraway CT, *et al.*** Cholera on a Gulf Coast oil rig. *New England Journal of Medicine* 1983;309:523–526.

272. **Gergatz SJ, McFarland LM**. Cholera on the Louisiana Gulf Coast: Historical notes and case report. *Journal of the Lousiana State Medical Society* 1989;141:29–34.

273. **Chen F, Evins GM, Cook WL, Almeida R, Hargrett-Bean N, *et al.*** Genetic diversity among toxigenic and nontoxigenic *Vibrio cholerae* O1 isolated from the Western Hemisphere. *Epidemiolgy and Infection* 1991;107:225–233.

274. **Cameron DN, Khambaty FM, Wachsmuth IK, Tauxe RV, Barrett TJ**. Molecular characterization of *Vibrio cholerae* O1 strains by pulsed-field gel electrophoresis. *Journal of Clinical Microbiology* 1994;32:1685–1690.

275. **Onifade TM, Hutchinson R, Zile KV, Bodager D, Baker R, *et al.*** Toxin producing *Vibrio cholerae* O75 outbreak, United States, March to April 2011. *Eurosurveillance* 2011;16:19870.

276. **Haley BJ, Choi SY, Grim CJ, Onifade TJ, Cinar HN, *et al.*** Genomic and phenotypic characterization of *Vibrio cholerae* non-O1 isolates from a US Gulf Coast cholera outbreak. *PLoS ONE* 2014;9:e86264.

277. **Centers for Disease Control and Prevention (CDC)**. Non-O1 and non-O139 infections | Sources of infection & risk factors | Cholera | CDC. https://www.cdc.gov/cholera/non-01-0139-infections.html (accessed 18 March 2018).

278. **Chowdhury G, Joshi S, Bhattacharya S, Sekar U, Birajdar B, *et al.*** Extraintestinal infections caused by non-toxigenic *Vibrio cholerae* non-O1/non-O139. *Frontiers in Microbiology* 2016;7:e00144.

279. **Dalsgaard A, Forslund A, Hesselbjerg A, Bruun B**. Clinical manifestations and characterization of extra-intestinal *Vibrio cholerae* non-O1, non-O139 infections in Denmark. *Clinical Microbiology and Infection* 2000;6:625–627.

280. **Farina C, Gnecchi F, Luzzi I, Vailati F**. *Vibrio cholerae* O2 as a cause of a skin lesion in a tourist returning from Tunisia. *Journal of Travel Medicine* 2000;7:92–94.

281. **Daniel D, Kumar S**. Rare strain of *Vibrio cholerae* septicemia in a patient with multiple myeloma. *Case Reports in Critical Care* 2015:596906.

282. **Kadkhoda K, Adam H, Gilmour MW, Hammond GW**. Nontoxigenic *Vibrio cholerae* septicemia in an immunocompromised patient. *Case Reports in Infectious Disease* 2012:698746.

283. **Garrine M, Mandomando I, Vubil D, Nhampossa T, Acacio S, *et al.*** Minimal genetic change in *Vibrio cholerae* in Mozambique over time: Multilocus variable number tandem repeat analysis and whole genome sequencing. *PLoS Neglected Tropical Diseases* 2017;11:e0005671.

284. **Sharma DP, Thomas C, Hall RH, Levine MM, Attridge SR**. Significance of toxin-coregulated pili as protective antigens of *Vibrio cholerae* in the infant mouse model. *Vaccine* 1989;7:451–456.

285. **Everiss KD, Hughes KJ, Kovach ME, Peterson KM**. The *Vibrio cholerae acfB* colonization determinant encodes an inner membrane protein that is related to a family of signal-transducing proteins. *Infection and Immunity* 1994;62:3289.

286. **Valiente E, Davies C, Mills DC, Getino M, Ritchie JM, *et al.*** *Vibrio cholerae* accessory colonisation factor AcfC: A chemotactic protein with a role in hyperinfectivity. *Scientific Reports* 2018;8:8390.

287. **Craig JP, Yamamoto K, Takeda Y, Miwatani T**. Production of cholera-like enterotoxin by a *Vibrio cholerae* non-O1 strain isolated from the environment. *Infection and Immunity* 1981;34:90–97.

288. **Ogawa A, Kato J, Watanabe H, Nair BG, Takeda T**. Cloning and nucleotide sequence of a heat-stable enterotoxin gene from *Vibrio cholerae* non-O1 isolated from a patient with traveler's diarrhea. *Infection and Immunity* 1990;58:3325–3329.

289. **Honda T, Arita M, Takeda T, Yoh M, Miwatani T**. Non-O1 *Vibrio cholerae* produces two newly identified toxins related to *Vibrio parahaemolyticus* haemolysin and *Escherichia coli* heat-stable enterotoxin. *The Lancet* 1985;2:163–164.

290. **Morris JG, Takeda T, Tall BD, Losonsky GA, Bhattacharya SK, *et al.*** Experimental non-O group 1 *Vibrio cholerae* gastroenteritis in humans. *Journal of Clinical Investigation* 1990;85:697–705.

291. **Dalsgaard A, Serichantalergs O, Shimada T, Sethabutr O, Echeverria P**. Prevalence of *Vibrio cholerae* with heat-stable enterotoxin (NAG-ST) and cholera toxin genes; restriction fragment length polymorphisms of NAG-ST genes among *V. cholerae* O serogroups from a major shrimp production area in Thailand. *Journal of Medical Microbiology* 1995;43:216–220.

292. **Hasan NA, Ceccarelli D, Grim CJ, Taviani E, Choi J, *et al.*** Distribution of virulence genes in clinical and environmental *Vibrio cholerae* strains in Bangladesh. *Applied and Environmental Microbiology* 2013;79:5782–5785.

293. **Mallard KE, Desmarchelier PM**. Detection of heat-stable enterotoxin genes among Australian *Vibrio cholerae* O1 strains. *FEMS Microbiology Letters* 1995;127:111–115.

294. **Takeda T, Peina Y, Ogawa A, Dohi S, Abe H, *et al*.** Detection of heat-stable enterotoxin in a cholera toxin gene-positive strain of *Vibrio cholerae* O1. *FEMS Microbiology Letters* 1991;80:23-27.

295. **Yuan P, Ogawa A, Ramamurthy T, Nair GB, Shimada T, *et al.*** *Vibrio mimicus* are the reservoirs of the heat-stable enterotoxin gene (*nag-st*) among species of the genus *Vibrio*. *World Journal of Microbiology and Biotechnology* 1994;10:59–63.

296. **Linhartová I, Bumba L, Mašín J, Basler M, Osička R, *et al.*** RTX proteins: A highly diverse family secreted by a common mechanism. *FEMS Microbiology Review* 2010;34:1076–1112.

297. **Lin W, Fullner KJ, Clayton R, Sexton JA, Rogers MB, *et al.*** Identification of a *Vibrio cholerae* RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. *Proceedings of the National Academy of Sciences of the United States of America* 1999;96:1071–1076.

298. **Satchell KJF**. Structure and function of MARTX toxins and other large repetitive RTX proteins. *Annual Review of Microbiology* 2011;65:71–90.

299. **Satchell KJF**. MARTX, multifunctional autoprocessing repeats-in-toxin toxins. *Infection and Immunity* 2007;75:5079–5084.

300. **Olivier V, Haines GK, Tan Y, Satchell KJF**. Hemolysin and the multifunctional autoprocessing RTX toxin are virulence factors during intestinal infection of mice with *Vibrio cholerae* El Tor O1 strains. *Infection and Immunity* 2007;75:5035–5042.

301. **Boardman BK, Fullner Satchell KJ**. *Vibrio cholerae* strains with mutations in an atypical type I secretion system accumulate RTX toxin intracellularly. *Journal of Bacteriology* 2004;186:8137–8143.

302. **Dolores J, Satchell KJF**. Analysis of *Vibrio cholerae* genome sequences reveals unique *rtxA* variants in environmental strains and an *rtxA*-null mutation in recent altered El Tor isolates. *mBio* 2013;4:e00624-12.

303. **Miller KA, Tomberlin KF, Dziejman M**. *Vibrio* variations on a type three theme. *Current Opinion in Microbiology* 2019;47:66–73.

304. **Mahmud J, Rashed SM, Islam T, Islam S, Watanabe H, *et al.*** Type three secretion system in non-toxigenic *Vibrio cholerae* O1, Mexico. *Journal of Medical Microbiology* 2014;63:1760–1762.

305. **Shin OS, Tam VC, Suzuki M, Ritchie JM, Bronson RT, *et al.*** Type III secretion is essential for the rapidly fatal diarrheal disease caused by non-O1, non-O139 *Vibrio cholerae*. *mBio* 2011;2:e00106-11.

306. **Garaizar J, Rementeria A, Porwollik S**. DNA microarray technology: A new tool for the epidemiological typing of bacterial pathogens? *FEMS Immunology and Medical Microbiology* 2006;47:178–189.

307. **Dutta A, Kundu JK, Chatterjee R, Chaudhuri K**. *In silico* comparative study of the genomic islands of *Vibrio cholerae* MJ1236 with those of Classical and El Tor N16961 strains of *Vibrio cholerae*. *FEMS Microbiology Letters* 2011;321:75–81.

308. **Didelot X, Pang B, Zhou Z, McCann A, Ni P, *et al.*** The role of China in the global spread of the current cholera pandemic. *PLoS Genetics* 2015;11:e1005072.

309. **Weill F-X, Domman D, Njamkepo E, Almesbahi AA, Naji M, *et al.*** Genomic insights into the 2016–2017 cholera epidemic in Yemen. *Nature* 2019;565:230–233.

310. **Hu D, Yin Z, Yuan C, Yang P, Qian C, *et al.*** Changing molecular epidemiology of *Vibrio cholerae* outbreaks in Shanghai, China. *mSystems* 2019;4:mSystems.00561-19.

311. **Shah MA, Mutreja A, Thomson N, Baker S, Parkhill J, *et al.*** Genomic epidemiology of *Vibrio cholerae* O1 associated with floods, Pakistan, 2010. *Emerging Infectious Diseases* 2014;20:13–20.

312. **Katz LS, Petkau A, Beaulaurier J, Tyler S, Antonova ES, *et al.*** Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *mBio* 2013;4:mBio.00398-13.

313. **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al.*** The sequence alignment/map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.

314. **Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, *et al.*** SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* 2012;19:cmb.2012.0021.

315. **Page AJ, De Silva N, Hunt M, Quail MA, Parkhill J, *et al.*** Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microbial Genomics* 2016;2:e000083.

316. **Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, *et al.*** Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* 2013;10:563–569.

317. **Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, *et al.*** Circlator: Automated circularization of genome assemblies using long sequencing reads. *Genome Biology* 2015;16:294.

318. **Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, *et al.*** Pilon: An integrated iool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* 2014;9:e112963.

319. **Wick RR, Judd LM, Gorrie CL, Holt KE**. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Computational Biology* 2017;13:e1005595.

320. **Wick RR, Schultz MB, Zobel J, Holt KE**. Bandage: Interactive visualization of *de novo* genome assemblies. *Bioinformatics* 2015;31:3350–3352.

321. **Wood DE, Salzberg SL**. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology* 2014;15:R46.

322. **Seemann T**. Prokka: Rapid prokaryotic genome annotation. *Bioinformatics* 2014;30:2068–2069.

323. **O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, *et al.*** Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Research* 2016;44:D733–D745.

324. **Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, *et al.*** Roary: Rapid large-scale prokaryote pan genome analysis. *Bioinformatics* 2015;31:3691–3693.

325. **Harris SR, Feil EJ, Holden MTG, Quail MA, Nickerson EK, *et al.*** Evolution of MRSA during hospital transmission and intercontinental spread. *Science* 2010;327:469–474.

326. **Allué-Guardia A, Echazarreta M, Koenig SSK, Klose KE, Eppinger M**. Closed genome sequence of *Vibrio cholerae* O1 El Tor Inaba strain A1552. *Genome Announcements* 2018;6:e00098-18.

327. **Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, *et al.*** Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Research* 2015;43:e15–e15.

328. **Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T**. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009;25:1972–1973.

329. **Page AJ, Taylor B, Delaney AJ, Soares J, Seemann T, *et al.*** SNP-sites: Rapid efficient extraction of SNPs from multi-FASTA alignments. *Microbial Genomics* 2016;2:e000056.

330. **Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ**. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution* 2015;32:268–274.

331. **Lewis PO**. A likelihood approach to estimating phylogeny from discrete morphological character data. *Systematic Biology* 2001;50:913–925.

332. **Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, *et al.*** New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biol*ogy 2010;59:307–321.

333. **Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS**. UFBoot2: Improving the ultrafast bootstrap approximation. *Molecular Biology and Evolution* 2018;35:518–522.

334. **Tonkin-Hill G, Lees JA, Bentley SD, Frost SDW, Corander J**. Fast hierarchical Bayesian analysis of population structure. *Nucleic Acids Research* 2019;47:5539–5549.

335. **Heller KA, Ghahramani Z**. Bayesian hierarchical clustering. In: *Proceedings of the 22nd International Conference on Machine Learning*. New York, NY, USA: ACM. pp. 297–304.

336. **Carver TJ, Rutherford KM, Berriman M, Rajandream M-A, Barrell BG, *et al.*** ACT: The Artemis comparison tool. *Bioinformatics* 2005;21:3422–3423.

337. **Sullivan MJ, Petty NK, Beatson SA**. Easyfig: A genome comparison visualizer. *Bioinformatics* 2011;27:1009–1010.

338. **Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ**. Basic local alignment search tool. *Journal of Molecular Biology* 1990;215:403–410.

339. **Carver T, Böhme U, Otto TD, Parkhill J, Berriman M**. BamView: Viewing mapped read alignment data in the context of the reference sequence. *Bioinformatics* 2010;26:676–677.

340. **Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S**. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* 2018;9:5114.

341. **Rambaut A, Lam TT, Max Carvalho L, Pybus OG**. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol* 2016;2:vew007.

342. **Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, *et al.*** ARIBA: Rapid antimicrobial resistance genotyping directly from sequencing reads. *Microbial Genomics* 2017;3:e000131.

343. **Zankari E, Hasman H, Cosentino S, Vestergaard M, Rasmussen S, *et al.*** Identification of acquired antimicrobial resistance genes. *Journal of Antimicrobial Chemotherapy* 2012;67:2640–2644.

344. **Carattoli A, Zankari E, García-Fernández A, Larsen MV, Lund O, *et al.*** *In silico* detection and typing of plasmids using PlasmidFinder and plasmid multilocus sequence typing. *Antimicrobial Agents and Chemotherapy* 2014;58:3895–3903.

345. **Grant JR, Arantes AS, Stothard P**. Comparing thousands of circular genomes using the CGView comparison tool. *BMC Genomics* 2012;13:202.

346. **R Core Team**. *R: A language and environment for statistical computing*. https://www.R-project.org/ (2018, accessed 20 June 2020).

347. **Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, *et al.*** Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology* 2011;7:msb.2011.75.

348. **Gouy M, Guindon S, Gascuel O**. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular Biology and Evolution* 2010;27:221–224.

349. **Bailey TL, Boden M, Buske FA, Frith M, Grant CE, *et al.*** MEME suite: Tools for motif discovery and searching. *Nucleic Acids Research* 2009;37:W202-208.

350. **McClure R, Balasubramanian D, Sun Y, Bobrovskyy M, Sumby P, *et al.*** Computational analysis of bacterial RNA-Seq data. *Nucleic Acids Research* 2013;41:e140.

351. **Tjaden B**. De novo assembly of bacterial transcriptomes from RNA-seq data. *Genome Biology* 2015;16:1–10.

352. **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 Research* 2016;44:W242–W245.

353. **Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, *et al.*** Phandango: An interactive viewer for bacterial population genomics. *Bioinformatics* 2018;34:292–293.

354. **Wickham H**. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York; 2016.

355. **Wickham H**. Reshaping data with the *reshape* package. *Journal of Statistical Software* 2007;21:1–20.

356. **Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, *et al.*** Artemis: Sequence visualization and annotation. *Bioinformatics* 2000;16:944–945.

357. **Carver T, Thomson N, Bleasby A, Berriman M, Parkhill J**. DNAPlotter: circular and linear interactive genome visualization. *Bioinformatics* 2009;25:119–120.

358. **Vella EE**. Cholera vaccines and the El Tor *Vibrio*. *British Medical Journal* 1963;1:1203–1207.

359. **Chang AC, Cohen SN**. Construction and characterization of amplifiable multicopy DNA cloning vehicles derived from the p15A cryptic miniplasmid. *Journal of Bacteriology* 1978;134:1141–1156.

360. **Norrander J, Kempe T, Messing J**. Construction of improved M13 vectors using oligodeoxynucleotide-directed mutagenesis. *Gene* 1983;26:101–106.

361. **Chowdhury F, Khan AI, Harris JB, LaRocque RC, Chowdhury MI, *et al.*** A comparison of clinical and immunologic features in children and older patients hospitalized with severe cholera in Bangladesh. *Pediatric Infectious Disease Journal* 2008;27:986–992.

362. **Reeves PR, Lan R**. Cholera in the 1990s. *British Medical Bulletin* 1998;54:611–623.

363. **Pan American Health Organisation (PAHO)**. Cholera situation in the Americas - Update. *Epidemiological Bulletin* 1992;13:11–12.

364. **Wilson MM, Chelala C**. Cholera is walking South. *Journal of the American Medical Association* 1994;272:1226–1227.

365. **Harrison A**. Cholera takes hold in northern Argentina. *United Press International*, 2 June 1992. https://www.upi.com/Archives/1992/02/06/Cholera-takes-hold-in-northern-Argentina/7518697352400/ (2 June 1992, accessed 12 March 2019).

366. **Petroni A, Corso A, Melano R, Cacace ML, Bru AM, *et al.*** Plasmidic extended-spectrum β-lactamases in *Vibrio cholerae* O1 El Tor isolates in Argentina. *Antimicrobial Agents and Chemotherapy* 2002;46:1462–1468.

367. **Pichel M, Rivas M, Chinen I, Martín F, Ibarra C, *et al.*** Genetic diversity of *Vibrio cholerae* O1 in Argentina and emergence of a new variant. *Journal of Clinical Microbiology* 2003;41:124–134.

368. **Castañeda NC, Pichel M, Orman B, Binsztein N, Roy PH, *et al.*** Genetic characterization of *Vibrio cholerae* isolates from Argentina by *V. cholerae* repeated sequences–polymerase chain reaction. *Diagnostic Microbiology and Infectious Disease* 2005;53:175–183.

369. **Pan American Health Organisation (PAHO)**. Cholera situation in the Americas, 1996. *Epidemiological Bulletin* 1997;18:5–7.

370. **Pan American Health Organisation (PAHO)**. Cholera in the Americas. *Epidemiological Bulletin* 1993;14:14.

371. **Pan American Health Organisation (PAHO)**. Cholera situation in the Americas. *Epidemiological Bulletin* 1994;15:13–16.

372. **Pan American Health Organisation (PAHO)**. Cholera in the Americas. *Epidemiological Bulletin* 1995;16:11–12.

373. **Eberhart-Phillips J, Besser RE, Tormey MP, Koo D, Feikin D, *et al.*** An outbreak of cholera from food served on an international aircraft. *Epidemiology and Infection* 1996;116:9–13.

374. **Centers for Disease Control and Prevention (CDC)**. Cholera associated with an international airline flight, 1992. *Morbidity and Mortality Weekly Report* 1992;41:134-135.

375. **Mydans S**. Cholera kills one and fells many on flight. *The New York Times*, 21 February 1992. https://www.nytimes.com/1992/02/21/us/cholera-kills-one-and-fells-many-on-flight.html (21 February 1992, accessed 12 March 2019).

376. **Lota L**. 5 People aboard Aerolineas Argentina flight 386 contract cholera. *Associated Press*, 20 February 1992. https://apnews.com/94bc1d2b18b7c4dc903d17f1db8a35cb (20 February 1992, accessed 12 March 2019).

377. **Nash NC**. Latin nations feud over cholera outbreak. *The New York Times*, 10 March 1992. https://www.nytimes.com/1992/03/10/world/latin-nations-feud-over-cholera-outbreak.html (10 March 1992, accessed 12 March 2019).

378. **Yildiz FH, Schoolnik GK**. Role of *rpoS* in stress survival and virulence of *Vibrio cholerae*. *Journal of Bacteriology* 1998;180:773–784.

379. **Blokesch M**. A quorum sensing-mediated switch contributes to natural transformation of *Vibrio cholerae*. *Mobile Genetic Elements* 2012;2:224–227.

380. **Mata L**. Cholera El Tor in Latin America, 1991-1993. *Annals of the New York Academy of Science* 1994;740:55–68.

381. **Wilson MM, Juliá CM, Chelala C**. How Argentina benefited from a cholera epidemic. *The Lancet* 1997;349:1375.

382. **Mouriño-Perez RR**. Oceanography and the seventh cholera pandemic. *Epidemiology* 1998;9:355–357.

383. **Mavian C, Paisie TK, Alam MT, Browne C, Rochars VMBD, *et al.*** Toxigenic *Vibrio cholerae* evolution and establishment of reservoirs in aquatic ecosystems. *Proceedings of the National Academy of Sciences of the United States of America*. Epub ahead of print 27 March 2020. DOI: 10.1073/pnas.1918763117.

384. **Martinez-Urtaza J, Trinanes J, Gonzalez-Escalona N, Baker-Austin C**. Is El Niño a long-distance corridor for waterborne disease? *Nature Microbiology* 2016;1:1–3.

385. **World Health Organization.** WHO | Weekly epidemiological record: cholera articles. *WHO*. http://www.who.int/cholera/statistics/en/ (accessed 18 June 2020).

386. **Pan American Health Organisation (PAHO)**. PAHO Regional Health Observatory - PHIP - Epidemic diseases - Cholera cases in the Americas since 1991. *PAHO Regional Health Observatory*. http://ais.paho.org/phip/viz/ed_colera_casesamericas.asp (accessed 18 June 2020).

387. **Rossi A, Galas M, Binztein N, Rivas M, Caffer MI, *et al.*** Unusual multiresistant *Vibrio cholerae* 01 El Tor in Argentina. *The Lancet* 1993;342:1172–1173.

388. **Price MN, Dehal PS, Arkin AP**. FastTree 2 – approximately maximum-likelihood trees for large alignments. *PLoS ONE* 2010;5:e9490.

389. **Fraga SG**. *Thesis*. Universidad de Buenos Aires; 2010.

390. **Matthey N, Drebes ND, Blokesch M**. Long-read-based genome sequences of pandemic and environmental *Vibrio cholerae* strains. *Microbiology Resource Announcements* 2018;7:MRA.01574-18.

391. **Eisenstark A**. Genetic diversity among offspring from archived *Salmonella enterica* ssp. *enterica* serovar Typhimurium (Demerec collection): In search of survival strategies. *Annual Review of Microbiology* 2010;64:277–292.

392. **Paul K, Ghosh A, Sengupta N, Chowdhury R**. Competitive growth advantage of nontoxigenic mutants in the stationary phase in archival cultures of pathogenic *Vibrio cholerae* strains. *Infection and Immunity* 2004;72:5478–5482.

393. **Carraro N, Sauvé M, Matteau D, Lauzon G, Rodrigue S, *et al.*** Development of pVCR94ΔX from *Vibrio cholerae*, a prototype for studying multidrug resistant IncA/C conjugative plasmids. *Frontiers in Microbiology* 2014;5:fmicb.2014.00044.

394. **Wang R, Liu H, Zhao X, Li J, Wan K**. IncA/C plasmids conferring high azithromycin resistance in *Vibrio cholerae*. *International Journal of Antimicrobial Agents* 2018;51:140–144.

395. **Harmer CJ, Hall RM**. The A to Z of A/C plasmids. *Plasmid* 2015;80:63–82.

396. **Wang H, Yang C, Sun Z, Zheng W, Zhang W, *et al.*** Genomic epidemiology of *Vibrio cholerae* reveals the regional and global spread of two epidemic non-toxigenic lineages. *PLoS Neglected Tropical Diseases* 2020;14:e0008046.

397. **Carpenter MR, Kalburge SS, Borowski JD, Peters MC, Colwell RR, *et al.*** CRISPR-Cas and contact-dependent secretion systems present on excisable pathogenicity islands with conserved recombination modules. *Journal of Bacteriology* 2017;199:e00842-16.

398. **Castillo D, Alvise PD, Xu R, Zhang F, Middelboe M, *et al.*** Comparative genome analyses of *Vibrio anguillarum* strains reveal a link with pathogenicity traits. *mSystems* 2017;2:mSystems.00001-17.

399. **Popovic T, Bopp C, Olsvik O, Wachsmuth K**. Epidemiologic application of a standardized ribotype scheme for *Vibrio cholerae* O1. *Journal of Clinical Microbiology* 1993;31:2474–2482.

400. **Wachsmuth IK, Evins GM, Fields PI, Olsvik Ø, Popovic T, *et al.*** The molecular epidemiology of cholera in Latin America. *Journal of Infectious Diseases* 1993;167:621–626.

401. **Pan American Health Organisation (PAHO)**. Cholera in the Americas - special report: Cholera. *Bulletin of PAHO* 1991;25:267–273.

402. **Guglielmetti P, Bartoloni A, Roselli M, Gamboa H, Antunez DJ, *et al.*** Population movements and cholera spread in Cordillera Province, Santa Cruz Department, Bolivia. *The Lancet* 1992;340:113.

403. **Evins GM, Cameron DN, Wells JG, Greene KD, Popovic T, *et al.*** The emerging diversity of the electrophoretic types of *Vibrio cholerae* in the Western Hemisphere. *Journal of Infectious Diseases* 1995;172:173–179.

404. **Dalsgaard A, Skov MN, Serichantalergs O, Echeverria P, Meza R, *et al.*** Molecular evolution of *Vibrio cholerae* O1 strains isolated in Lima, Peru, from 1991 to 1995. *Journal of Clinical Microbiology* 1997;35:1151–1156.

405. **Pan American Health Organisation (PAHO)**. Cholera situation in the Americas. An update. *Epidemiological Bulletin* 1991;12:1–4.

406. **Iqbal N, Guérout A-M, Krin E, Roux FL, Mazel D**. Comprehensive functional analysis of the 18 *Vibrio cholerae* N16961 toxin-antitoxin systems substantiates their role in stabilizing the superintegron. *Journal of Bacteriology* 2015;197:2150–2159.

407. **Chain PSG, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, *et al.*** Genome project standards in a new era of sequencing. *Science* 2009;326:236-237.

408. **H. M. Government**. Anti-Terrorism, Crime and Security Act 2001 (Modification) Order 2007 (2007/929). 2007.

409. **Dorman MJ, Domman D, Uddin MI, Sharmin S, Afrad MH, *et al.*** High quality reference genomes for toxigenic and non-toxigenic *Vibrio cholerae* serogroup O139. *Scientific Reports* 2019;9:5865.

410. **Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, *et al.*** Canu: Scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Research* 2017;gr.215087.116.

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

412. **Mukhopadhyay AK, Basu A, Garg P, Bag PK, Ghosh A, *et al.*** Molecular epidemiology of reemergent *Vibrio cholerae* O139 Bengal in India. *Journal of Clinical Microbiology* 1998;36:2149–2152.

413. **Kimsey HH, Nair GB, Ghosh A, Waldor MK**. Diverse CTXΦs and evolution of new pathogenic *Vibrio cholerae*. *The Lancet* 1998;352:457–458.

414. **Sharma C, Maiti S, Mukhopadhyay AK, Basu A, Basu I, *et al.*** Unique organization of the CTX genetic element in *Vibrio cholerae* O139 strains which reemerged in Calcutta, India, in September 1996. *Journal of Clinical Microbiology* 1997;35:3348–3350.

415. **Bhuiyan NA, Nusrin S, Alam M, Morita M, Watanabe H, *et al.*** Changing genotypes of cholera toxin (CT) of *Vibrio cholerae* O139 in Bangladesh and description of three new CT genotypes. *FEMS Immunology & Medical Microbiology* 2009;57:136–141.

416. **Kim EJ, Lee CH, Nair GB, Kim DW**. Whole-genome sequence comparisons reveal the evolution of *Vibrio cholerae* O1. *Trends in Microbiology* 2015;23:479–489.

417. **Davis BM, Kimsey HH, Chang W, Waldor MK**. The *Vibrio cholerae* O139 Calcutta bacteriophage CTXφ is infectious and encodes a novel repressor. *Journal of Bacteriology* 1999;181:6779–6787.

418. **Klinzing DC, Choi SY, Hasan NA, Matias RR, Tayag E, *et al.*** Hybrid *Vibrio cholerae* El Tor lacking SXT identified as the cause of a cholera outbreak in the Philippines. *mBio* 2015;6:e00047-15.

419. **Grim CJ, Choi J, Chun J, Jeon Y-S, Taviani E, *et al.*** Occurrence of the *Vibrio cholerae* seventh pandemic VSP-I island and a new variant. *OMICS: A Journal of Integrative Biology* 2010;14:1–7.

420. **Rowe-Magnus DA, Guerout A-M, Mazel D**. Bacterial resistance evolution by recruitment of super-integron gene cassettes. *Molecular Microbiology* 2002;43:1657–1669.

421. **Singh A**. Cholera and its prevention. *The Indian Medical Gazette* 1917;52:31–32.

422. **Bennett VB**. Cholera. *The Indian Medical Gazette* 1917;52:363.

423. **Mitchell TJ, Smith GM**. *Medical services; casualties and medical statistics of the Great War*. London. http://hdl.handle.net/2027/uc1.$b744277 (1931).

424. **Dorman MJ, Kane L, Domman D, Turnbull JD, Cormie C, *et al.*** The history, genome and biology of NCTC 30: a non-pandemic *Vibrio cholerae* isolate from World War One. *Proceedings of the Royal Society B: Biological Sciences* 2019;286:20182025.

425. **Dorman MJ, Thomson NR**. 'Community evolution' – laboratory strains and pedigrees in the age of genomics. *Microbiology* 2020;166:000869.

426. **Klose KE, Mekalanos JJ**. Distinct roles of an alternative sigma factor during both free-swimming and colonizing phases of the *Vibrio cholerae* pathogenic cycle. *Molecular Microbiology* 1998;28:501–520.

427. **Prouty MG, Correa NE, Klose KE**. The novel σ^{54}- and σ^{28}-dependent flagellar gene transcription hierarchy of *Vibrio cholerae*. *Molecular Microbiology* 2001;39:1595–1609.

428. **Syed KA, Beyhan S, Correa N, Queen J, Liu J, *et al.*** The *Vibrio cholerae* flagellar regulatory hierarchy controls expression of virulence factors. *Journal of Bacteriology* 2009;191:6555–6570.

429. **Cowan ST**. The National Collection of Type Cultures. *The Lancet* 1950;255:82–83.

430. **Correa NE, Lauriano CM, McGee R, Klose KE**. Phosphorylation of the flagellar regulatory protein FlrC is necessary for *Vibrio cholerae* motility and enhanced colonization. *Molecular Microbiology* 2000;35:743–755.

431. **Feng L, Reeves PR, Lan R, Ren Y, Gao C, *et al.*** A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS ONE* 2009;3:e4053.

432. **Institute of Medicine (US) Committee on Issues and Priorities for New Vaccine Development**. *The prospects for immunizing against* Vibrio cholerae. National Academies Press (US). https://www.ncbi.nlm.nih.gov/books/NBK219074/ (1986, accessed 19 June 2020).

433. **Fleming A**. On the antibacterial action of cultures of a *Penicillium*, with special reference to their use in the isolation of *B. influenzæ*. *British Journal of Experimental Pathology* 1929;10:226–236.

434. **World Health Organization**. WHO | Prevention and control of cholera outbreaks: WHO policy and recommendations.

http://www.who.int/cholera/prevention_control/recommendations/en/index4.html (2018, accessed 31 March 2018).

435. **Miller VL, DiRita VJ, Mekalanos JJ**. Identification of *toxS*, a regulatory gene whose product enhances *toxR*-mediated activation of the cholera toxin promoter. *Journal of Bacteriology* 1989;171:1288–1293.

436. **Bhattacharya MK, Bhattacharya SK, Garg S, Saha PradipK, Dutta D, *et al.*** Outbreak of *Vibrio cholerae* non-01 in India and Bangladesh. *The Lancet* 1993;341:1346–1347.

437. **Hammer BK, Bassler BL**. Regulatory small RNAs circumvent the conventional quorum sensing pathway in pandemic *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 2007;104:11145–11149.

438. **Joelsson A, Liu Z, Zhu J**. Genetic and phenotypic diversity of quorum-sensing systems in clinical and environmental isolates of *Vibrio cholerae*. *Infection and Immunity* 2006;74:1141–1147.

439. **Greig DR, Schaefer U, Octavia S, Hunter E, Chattaway MA, *et al.*** Evaluation of whole-genome sequencing for identification and typing of *Vibrio cholerae*. *Journal of Clinical Microbiology* 2018;56: e00831-18.

440. **Okada K, Na-Ubol M, Natakuathung W, Roobthaisong A, Maruyama F, *et al.*** Comparative genomic characterization of a Thailand-Myanmar isolate, MS6, of *Vibrio cholerae* O1 El Tor, which is phylogenetically related to a 'US Gulf Coast' clone. *PLoS ONE* 2014;9:e98120.

441. **Nonaka L, Suzuki S**. New Mg^{2+}-dependent oxytetracycline resistance determinant *tet34* in *Vibrio* isolates from marine fish intestinal contents. *Antimicrobial Agents and Chemotherapy* 2002;46:1550–1552.

442. **Burnet FM**. The initiation of cellular infection by influenza and related viruses. *The Lancet* 1948;251:7–11.

443. **Judicial Commission of the International Committee on Bacteriological Nomenclature**. Opinion 31. Conservation of *Vibrio* Pacini 1854 as a bacterial generic

name, conservation of *Vibrio cholerae* Pacini 1854 as the nomenclatural type species of the bacterial genus *Vibrio*, and designation of neotype strain of *Vibrio cholerae* Pacini. *International Journal of Systematic and Evolutionary Microbiology,* 1965;15:185–186.

444. **Hugh R**. A comparison of the proposed neotype strain and 258 isolates of *Vibrio cholerae* Pacini. *International Journal of Systematic and Evolutionary Microbiology,* 1965;15:13–24.

445. **Burrows W, Mather AN, Elliott ME, Wagner SM**. Studies on immunity to Asiatic cholera I. Introduction. *Journal of Infectious Diseases* 1946;79:159–167.

446. **Mukherjee B, De SN**. Observations on some newer differential features of the El Tor vibrio and *Vibrio cholerae*. *The Journal of Pathology and Bacteriology* 1966;91:256–263.

447. **Doorenbos W, Kop J**. El Tor Vibrio in chloramphenicol estimation. *The Lancet* 1951;257:691.

448. **Zhang R, Wang Y, Leung PC, Gu J-D**. pVC, a small cryptic plasmid from the environmental isolate of *Vibrio cholerae* MP-1. *The Journal of Microbiology* 2007;45:193–198.

449. **Lilly J, Camps M**. Mechanisms of theta plasmid replication. *Microbiology Spectrum* 2015;3: PLAS-0029-2014.

450. **Newland JW, Voll MJ, McNicol LA**. Serology and plasmid carriage in *Vibrio cholerae*. *Canadian Journal of Microbiology* 1984;30:1149–1156.

451. **McDevitt S**. Methyl red and Voges-Proskauer test protocols. *ASMScience Laboratory Protocols*. https://www.asmscience.org/content/education/protocol/protocol.3204 (2009, accessed 25 May 2020).

452. **Yoon SS, Mekalanos JJ**. 2,3-Butanediol synthesis and the emergence of the *Vibrio cholerae* El Tor biotype. *Infection and Immunity* 2006;74:6547–6556.

453. **Kovacikova G, Lin W, Skorupski K**. Dual regulation of genes involved in acetoin biosynthesis and motility/biofilm formation by the virulence activator AphA and the

acetate-responsive LysR-type regulator AlsR in *Vibrio cholerae*. *Molecular Microbiology* 2005;57:420–433.

454. **Oh YT, Lee K-M, Bari W, Raskin DM, Yoon SS**. (p)ppGpp, a small nucleotide regulator, directs the metabolic fate of glucose in *Vibrio cholerae*. *Journal of Biological Chemistry* 2015;290:13178–13190.

455. **Kovacikova G, Skorupski K**. Regulation of virulence gene expression in *Vibrio cholerae* by quorum sensing: HapR functions at the *aphA* promoter. *Molecular Microbiology* 2002;46:1135–1147.

456. **Iwanaga M, Yamamoto K**. New medium for the production of cholera toxin by *Vibrio cholerae* O1 biotype El Tor. *Journal of Clinical Microbiology* 1985;22:405–408.

457. **Iwanaga M, Yamamoto K, Naomi H, Ichinose Y, Nakasone N, *et al.*** Culture conditions for stimulating cholera toxin production by *Vibrio cholerae* O1 El Tor. *Microbiology and Immunology* 1986;30:1075–1083.

458. **Kovacikova G, Skorupski K**. A *Vibrio cholerae* LysR homolog, AphB, cooperates with AphA at the *tcpPH* promoter to activate expression of the ToxR virulence cascade. *Journal of Bacteriology* 1999;181:4250–4256.

459. **Kovacikova G, Skorupski K**. Binding site requirements of the virulence gene regulator AphB: differential affinities for the *Vibrio cholerae* classical and El Tor *tcpPH* promoters. *Molecular Microbiology* 2002;44:533–547.

460. **Kovacikova G, Skorupski K**. Differential activation of the *tcpPH* Promoter by AphB determines biotype specificity of virulence gene expression in *Vibrio cholerae*. *Journal of Bacteriology* 2000;182:3228–3238.

461. **Hankins JV, Madsen JA, Giles DK, Brodbelt JS, Trent MS**. Amino acid addition to *Vibrio cholerae* LPS establishes a link between surface remodeling in Gram-positive and Gram-negative bacteria. *Proceedings of the National Academy of Sciences of the United States of America* 2012;109:8722–8727.

462. **Hall RH, Drasar BS**. *Vibrio cholerae* HlyA hemolysin is processed by proteolysis. *Infection and Immunity* 1990;58:3375–3379.

463. **Yamamoto K, Ichinose Y, Shinagawa H, Makino K, Nakata A, *et al.*** Two-step processing for activation of the cytolysin/hemolysin of *Vibrio cholerae* O1 biotype El Tor: Nucleotide sequence of the structural gene (*hlyA*) and characterization of the processed products. *Infection and Immunity* 1990;58:4106–4116.

464. **Alm RA, Mayrhofer G, Kotlarski I, Manning PA**. Amino-terminal domain of the El Tor haemolysin of *Vibrio cholerae* O1 is expressed in classical strains and is cytotoxic. *Vaccine* 1991;9:588–594.

465. **Herrera CM, Crofts AA, Henderson JC, Pingali SC, Davies BW, *et al.*** The *Vibrio cholerae* VprA-VprB two-component system controls virulence through endotoxin modification. *mBio* 2014;5:e02283-14.

466. **Lee D, Kim EJ, Baek Y, Lee J, Yoon Y, *et al.*** Alterations in glucose metabolism in *Vibrio cholerae* serogroup O1 El Tor biotype strains. *Scientific Reports* 2020;10:308.

467. **Liang K, Islam MT, Hussain N, Winkjer NS, Im MS, *et al.*** Draft genome sequences of eight *Vibrio* sp. clinical isolates from across the United States that form a basal sister clade to *Vibrio cholerae*. *Microbiology Resource Announcements* 2019;8:MRA.01473-18.

468. **Xu Q, Dziejman M, Mekalanos JJ**. Determination of the transcriptome of *Vibrio cholerae* during intraintestinal growth and midexponential phase *in vitro*. *Proceedings of the National Academy of Sciences of the United States of America* 2003;100:1286–1291.

469. **Merrell DS, Butler SM, Qadri F, Dolganov NA, Alam A, *et al.*** Host-induced epidemic spread of the cholera bacterium. *Nature* 2002;417:642–645.

470. **Beyhan S, Tischler AD, Camilli A, Yildiz FH**. Differences in gene expression between the Classical and El Tor biotypes of *Vibrio cholerae* O1. *Infection and Immunity* 2006;74:3633–3642.

471. **Mandlik A, Livny J, Robins WP, Ritchie JM, Mekalanos JJ, *et al.*** RNA-seq-based monitoring of infection-linked changes in *Vibrio cholerae* gene expression. *Cell Host and Microbe* 2011;10:165–174.

472. **Krin E, Pierlé SA, Sismeiro O, Jagla B, Dillies M-A, *et al.*** Expansion of the SOS regulon of *Vibrio cholerae* through extensive transcriptome analysis and experimental validation. *BMC Genomics* 2018;19:373.

473. **Zhang Z, Chen G, Hu J, Hussain W, Fan F, *et al.*** Mr.Vc: A database of microarray and RNA-seq of *Vibrio cholerae*. *Database* 2019;2019:baz069.

474. **DuPai CD, Wilke CO, Davies BW**. A comprehensive coexpression network analysis in *Vibrio cholerae*. *mSystems*;5. Epub ahead of print 25 August 2020. DOI: 10.1128/mSystems.00550-20.

475. **Papenfort K, Förstner KU, Cong J-P, Sharma CM, Bassler BL**. Differential RNA-seq of *Vibrio cholerae* identifies the VqmR small RNA as a regulator of biofilm formation. *Proceedings of the National Academy of Sciences of the United States of America* 2015;112:E766–E775.

476. **Miller VL, Mekalanos JJ**. A novel suicide vector and its use in construction of insertion mutations: Osmoregulation of outer membrane proteins and virulence determinants in *Vibrio cholerae* requires *toxR*. *Journal of Bacteriology* 1988;170:2575–2583.

477. **Peterson KM, Mekalanos JJ**. Characterization of the *Vibrio cholerae* ToxR regulon: Identification of novel genes involved in intestinal colonization. *Infection and Immunity* 1988;56:2822–2829.

478. **Skorupski K, Taylor RK**. Cyclic AMP and its receptor protein negatively regulate the coordinate expression of cholera toxin and toxin-coregulated pilus in *Vibrio cholerae*. *Proceedings of the National Academy of Sciences of the United States of America* 1997;94:265–270.

479. **Rashid M, Rashed SM, Islam T, Johura F-T, Watanabe H, *et al.*** CtxB1 outcompetes CtxB7 in *Vibrio cholerae* O1, Bangladesh. *Journal of Medical Microbiology* 2016;65:101–103.

480. **Basler M, Pilhofer M, Henderson GP, Jensen GJ, Mekalanos JJ**. Type VI secretion requires a dynamic contractile phage tail-like structure. *Nature* 2012;483:182–186.

481. **Basler M, Ho BT, Mekalanos JJ**. Tit-for-tat: Type VI secretion system counterattack during bacterial cell-cell interactions. *Cell* 2013;152:884–894.

482. **Peng Y, Wang X, Shou J, Zong B, Zhang Y, *et al.*** Roles of Hcp family proteins in the pathogenesis of the porcine extraintestinal pathogenic *Escherichia coli* type VI secretion system. *Scientific Reports* 2016;6:26816.

483. **Silverman JM, Agnello DM, Zheng H, Andrews BT, Li M, *et al.*** Haemolysin coregulated protein is an exported receptor and chaperone of type VI secretion substrates. *Molecular Cell* 2013;51:584–593.

484. **Metzger LC, Stutzmann S, Scrignari T, Van der Henst C, Matthey N, *et al.*** Independent regulation of type VI secretion in *Vibrio cholerae* by TfoX and TfoY. *Cell Reports* 2016;15:951–958.

485. **Dong TG, Mekalanos JJ**. Characterization of the RpoN regulon reveals differential regulation of T6SS and new flagellar operons in *Vibrio cholerae* O37 strain V52. *Nucleic Acids Research* 2012;40:7766–7775.

486. **Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, *et al.*** Identification of a conserved bacterial protein secretion system in *Vibrio cholerae* using the *Dictyostelium* host model system. *Proceedings of the National Academy of Sciences of the United States of America* 2006;103:1528–1533.

487. **Kitaoka M, Miyata ST, Brooks TM, Unterweger D, Pukatzki S**. VasH is a transcriptional regulator of the type VI secretion system functional in endemic and pandemic *Vibrio cholerae*. *Journal of Bacteriology* 2011;193:6471–6482.

488. **Unterweger D, Kitaoka M, Miyata ST, Bachmann V, Brooks TM, *et al.*** Constitutive type VI secretion system expression gives *Vibrio cholerae* intra- and interspecific competitive advantages. *PLoS ONE* 2012;7:e48320.

489. **Miyata ST, Kitaoka M, Wieteska L, Frech C, Chen N, *et al.*** The *Vibrio cholerae* type VI secretion system: Evaluating its role in the human disease cholera. *Frontiers in Microbiology* 2010;1:fmicb.2010.00117.

490. **Hersch SJ, Watanabe N, Stietz MS, Manera K, Kamal F, *et al.*** Envelope stress responses defend against type six secretion system attacks independently of immunity proteins. *Nature Microbiology* 2020;5:706–714.

Appendix 1

List of *V. cholerae* sequenced for this PhD research

Sequencing data produced as part of this PhD have been, or will be, released into the European Nucleotide Archive under accession numbers PRJEB14661, ERP110583, and ERP118963.

Name	Serogroup	Serotype	Source	Isolation date	Country
CCBT0190	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0192	O1	Ogawa	Environmental	05/05/1992	Argentina
CCBT0187	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0194	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0264	O1	Inaba	Clinical	05/05/1992	Argentina
CCBT0258	O1	Ogawa	Clinical	20/06/1997	Argentina
CCBT0243	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0253	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0212	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0247	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0245	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0246	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0256	O1	Ogawa	Clinical	24/02/1992	Argentina
CCBT0235	O1	Ogawa	Clinical	01/04/1992	Argentina
CCBT0262	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0231	O1	Ogawa	Clinical	10/02/1992	Argentina
CCBT0232	O1	Ogawa	Clinical	12/03/1992	Argentina
CCBT0255	O1	Ogawa	Clinical	01/04/1992	Argentina
CCBT0172	O1	Ogawa	Clinical	10/02/1992	Argentina
CCBT0249	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0233	O1	Ogawa	Clinical	01/04/1992	Argentina
CCBT0062	O1	Ogawa	Clinical	24/10/1992	Argentina
CCBT0183	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0189	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0070	O1	Inaba	Clinical	19/10/1992	Argentina
CCBT0184	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0071	O1	/	Clinical	24/10/1992	Argentina
CCBT0067	O1	Ogawa	Clinical	19/10/1992	Argentina
CCBT0257	O1	Ogawa	Clinical	01/04/1992	Argentina
CCBT0064	O1	Ogawa	Clinical	01/09/1992	Argentina
CCBT0345	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0479	O1	Ogawa	Clinical	03/03/1993	Argentina
CCBT0097	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0471	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0295	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0379	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0400	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0450	O1	Ogawa	Clinical	12/02/1993	Argentina

CCBT0317	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0315	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0478	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0357	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0480	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0299	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0477	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0390	/	/	Clinical	27/01/1993	Argentina
CCBT0166	O1	Ogawa	Clinical	22/12/1992	Argentina
CCBT0224	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0488	O1	Ogawa	Environmental	20/01/1993	Argentina
CCBT0340	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0203	O1	Ogawa	Clinical	26/03/1993	Argentina
CCBT0461	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0120	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0396	O1	Ogawa	Clinical	15/02/1993	Argentina
CCBT0466	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0098	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0202	O1	Ogawa	Clinical	15/03/1993	Argentina
CCBT0341	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0122	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0123	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0276	O1	Ogawa	Clinical	08/03/1993	Argentina
CCBT0418	O1	Ogawa	Clinical	10/02/1993	Argentina
CCBT0439	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0388	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0381	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0473	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0428	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0423	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0424	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0285	O1	/	Clinical	22/01/1993	Argentina
CCBT0351	O1	Ogawa	Clinical	05/02/1993	Argentina
CCBT0132	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0356	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0413	/	/	Clinical	27/01/1993	Argentina
CCBT0175	O1	Ogawa	Clinical	30/03/1993	Argentina
CCBT0430	O1	Ogawa	Clinical	05/02/1993	Argentina
CCBT0292	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0394	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0436	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0318	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0414	/	/	Clinical	27/01/1993	Argentina

CCBT0337	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0353	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0116	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0442	O1	Ogawa	Clinical	08/03/1993	Argentina
CCBT0339	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0213	O1	Ogawa	Clinical	30/03/1993	Argentina
CCBT0425	O1	Inaba	Clinical	05/02/1993	Argentina
CCBT0481	O1	Inaba	Clinical	05/02/1993	Argentina
CCBT0326	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0144	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0459	O1	Ogawa	Clinical	08/03/1993	Argentina
CCBT0308	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0352	O1	Ogawa	Clinical	08/03/1993	Argentina
CCBT0321	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0446	O1	Ogawa	Clinical	05/02/1993	Argentina
CCBT0261	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0327	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0291	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0226	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0490	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0393	O1	Ogawa	Clinical	25/10/1993	Argentina
CCBT0322	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0294	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0266	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0323	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0173	O1	Inaba	Clinical	17/02/1992	Argentina
CCBT0306	O1	Inaba	Clinical	02/02/1993	Argentina
CCBT0447	O1	Inaba	Clinical	20/02/1996	Argentina
CCBT0110	O1	Ogawa	Clinical	01/12/1992	Argentina
CCBT0432	O1	Ogawa	Clinical	05/02/1993	Argentina
CCBT0105	O1	Ogawa	Clinical	28/12/1992	Argentina
CCBT0163	O1	Ogawa	Clinical	28/12/1992	Argentina
CCBT0108	O1	Ogawa	Clinical	25/11/1992	Argentina
CCBT0333	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0138	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0376	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0236	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0278	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0456	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0457	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0313	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0148	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0465	O1	Ogawa	Clinical	12/02/1993	Argentina

CCBT0330	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0154	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0209	O1	Ogawa	Clinical	30/03/1993	Argentina
CCBT0085	O1	Inaba	Clinical	09/04/1996	Argentina
CCBT0050	O1	Inaba	Clinical	24/04/1996	Argentina
CCBT0089	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0055	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0088	O1	Inaba	Clinical	30/05/1996	Argentina
CCBT0052	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0095	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0047	O1	Inaba	Clinical	06/06/1996	Argentina
CCBT0057	O1	Inaba	Clinical	06/06/1996	Argentina
CCBT0083	O1	Inaba	Clinical	06/06/1996	Argentina
CCBT0051	O1	Inaba	Clinical	28/05/1996	Argentina
CCBT0059	O1	Inaba	Clinical	28/02/1996	Argentina
CCBT0041	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0048	O1	Inaba	Clinical	28/05/1996	Argentina
CCBT0046	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0075	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0487	O1	Ogawa	Environmental	17/02/1994	Argentina
CCBT0486	O1	Ogawa	Clinical	15/02/1994	Argentina
CCBT0003	O1	Ogawa	Clinical	30/12/1996	Argentina
CCBT0049	O1	Ogawa	Clinical	03/06/1996	Argentina
CCBT0091	O1	Ogawa	Clinical	28/05/1996	Argentina
CCBT0056	O1	Ogawa	Clinical	03/06/1996	Argentina
CCBT0092	O1	Ogawa	Clinical	03/06/1996	Argentina
CCBT0035	O1	Ogawa	Clinical	06/01/1997	Argentina
CCBT0017	O1	Ogawa	Clinical	06/01/1997	Argentina
CCBT0019	O1	Ogawa	Clinical	06/01/1997	Argentina
CCBT0010	O1	Inaba	Clinical	10/02/1997	Argentina
CCBT0022	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0039	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0040	O1	Ogawa	Clinical	23/01/1997	Argentina
CCBT0023	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0029	O1	Ogawa	Clinical	06/01/1997	Argentina
CCBT0369	O1	Ogawa	Clinical	21/01/1997	Argentina
CCBT0014	O1	Ogawa	Clinical	13/01/1997	Argentina
CCBT0380	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0028	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0080	O1	Inaba	Clinical	03/06/1996	Argentina
CCBT0372	O1	Ogawa	Clinical	21/01/1997	Argentina
CCBT0373	/	/	/	/	Argentina
CCBT0012	O1	Ogawa	Clinical	08/11/1996	Argentina

CCBT0009	O1	Ogawa	Clinical	13/03/1997	Argentina
CCBT0037	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0363	O1	Ogawa	Clinical	21/01/1997	Argentina
CCBT0002	O1	Ogawa	Clinical	21/01/1997	Argentina
CCBT0006	O1	Ogawa	Clinical	13/01/1997	Argentina
CCBT0032	O1	Ogawa	Clinical	23/01/1997	Argentina
CCBT0031	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0013	O1	Ogawa	Clinical	23/01/1997	Argentina
CCBT0368	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0364	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0015	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0021	/	/	/	/	Argentina
CCBT0367	O1	Ogawa	Clinical	23/01/1997	Argentina
CCBT0374	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0034	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0016	O1	Ogawa	Clinical	08/11/1996	Argentina
CCBT0024	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0030	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0499	/	/	/	/	Argentina
CCBT0365	O1	Ogawa	Clinical	08/11/1996	Argentina
CCBT0366	O1	Ogawa	Clinical	08/11/1996	Argentina
CCBT0371	O1	Ogawa	Clinical	08/11/1996	Argentina
CCBT0001	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0038	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0025	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0036	O1	Ogawa	Clinical	07/02/1997	Argentina
CCBT0005	O1	Ogawa	Clinical	23/01/1997	Argentina
CCBT0011	/	/	/	/	Argentina
CCBT0004	O1	Ogawa	Clinical	23/01/1997	Argentina
CCBT0027	O1	Ogawa	Clinical	23/01/1997	Argentina
CCBT0498	O1	Inaba	Clinical	04/03/1996	Argentina
CCBT0100	O1	Ogawa	Clinical	13/11/1992	Argentina
CCBT0103	O1	Ogawa	Clinical	25/11/1992	Argentina
CCBT0111	O1	Ogawa	Clinical	25/11/1992	Argentina
CCBT0104	O1	Ogawa	Clinical	25/11/1992	Argentina
CCBT0114	O1	Ogawa	Clinical	25/11/1992	Argentina
CCBT0331	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0107	O1	Ogawa	Clinical	25/11/1992	Argentina
CCBT0409	O1	Ogawa	Clinical	22/02/1993	Argentina
CCBT0159	O1	Ogawa	Clinical	28/12/1992	Argentina
CCBT0164	O1	Ogawa	Clinical	22/12/1992	Argentina
CCBT0389	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0272	O1	Ogawa	Clinical	20/01/1993	Argentina

CCBT0199	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0312	/	/	Clinical	28/01/1993	Argentina
CCBT0489	O1	Ogawa	Environmental	17/01/1994	Argentina
CCBT0161	non-O1	/	Environmental	01/12/1992	Argentina
CCBT0344	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0251	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0470	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0463	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0415	O1	Ogawa	Clinical	22/02/1993	Argentina
CCBT0483	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0283	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0426	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0069	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0416	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0441	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0347	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0410	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0305	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0375	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0150	non-O1	/	Environmental	29/12/1992	Argentina
CCBT0383	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0217	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0128	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0421	O1	Ogawa	Clinical	22/02/1993	Argentina
CCBT0109	O1	Ogawa	Clinical	22/12/1992	Argentina
CCBT0259	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0160	O1	Ogawa	Clinical	28/12/1992	Argentina
CCBT0165	O1	Ogawa	Clinical	22/12/1992	Argentina
CCBT0384	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0408	O1	Ogawa	Clinical	22/02/1993	Argentina
CCBT0336	O1	Ogawa	Clinical	27/01/1993	Argentina
CCBT0482	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0358	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0434	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0121	O1	Ogawa	Clinical	06/01/1993	Argentina
CCBT0127	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0143	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0286	non-O1	/	Clinical	18/03/1996	Argentina
CCBT0152	O1	Ogawa	Environmental	19/01/1993	Argentina
CCBT0147	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0124	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0125	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0359	O1	Ogawa	Clinical	02/03/1993	Argentina

CCBT0469	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0401	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0449	O1	Ogawa	Clinical	02/03/1993	Argentina
VC005	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0454	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0329	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0451	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0134	O1	Ogawa	Environmental	19/01/1993	Argentina
CCBT0476	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0241	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0239	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0185	O1	Ogawa	Environmental	21/04/1992	Argentina
CCBT0349	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0182	O1	Ogawa	Clinical	26/03/1993	Argentina
CCBT0370	O1	Ogawa	Environmental	10/05/1993	Argentina
CCBT0468	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0265	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0474	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0279	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0435	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0271	O1	Ogawa	Clinical	21/04/1992	Argentina
VC007	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0201	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0215	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0270	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0404	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0126	O1	Ogawa	Environmental	19/01/1993	Argentina
CCBT0142	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0284	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0429	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0440	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0484	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0200	O1	Ogawa	Environmental	05/05/1992	Argentina
CCBT0302	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0398	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0392	O1	Inaba	Clinical	18/02/1993	Argentina
CCBT0288	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0275	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0386	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0198	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0303	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0492	O1	Ogawa	Environmental	10/05/1993	Argentina
CCBT0346	O1	Ogawa	Clinical	12/02/1993	Argentina

CCBT0153	O1	Ogawa	Environmental	19/01/1993	Argentina
CCBT0467	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0422	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0168	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0420	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0324	O1	Ogawa	Clinical	03/02/1993	Argentina
CCBT0133	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0335	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0139	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0207	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0417	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0301	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0156	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0222	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0412	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0065	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0225	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0287	O1	Ogawa	Clinical	22/01/1993	Argentina
VC011	non-O1	/	Clinical	21/04/1992	Argentina
CCBT0178	O1	Ogawa	Clinical	24/02/1992	Argentina
CCBT0277	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0068	O1	Ogawa	Clinical	20/05/1992	Argentina
CCBT0438	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0382	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0157	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0197	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0348	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0460	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0350	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0296	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0355	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0458	O1	Ogawa	Clinical	22/02/1993	Argentina
CCBT0234	O1	Ogawa	Clinical	01/04/1992	Argentina
CCBT0297	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0290	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0195	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0137	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0307	O1	/	Clinical	22/01/1993	Argentina
CCBT0485	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0167	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0211	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0402	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0186	O1	Ogawa	Clinical	26/03/1993	Argentina

CCBT0158	O1	Ogawa	Clinical	28/12/1992	Argentina
CCBT0452	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0181	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0238	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0268	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0250	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0196	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0169	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0248	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0170	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0433	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0300	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0332	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0455	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0310	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0437	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0204	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0237	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0260	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0360	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0273	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0462	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0338	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0378	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0228	O1	Ogawa	Clinical	26/03/1993	Argentina
CCBT0431	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0399	O1	Ogawa	Clinical	05/02/1993	Argentina
CCBT0171	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0191	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0267	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0188	O1	Ogawa	Environmental	05/05/1992	Argentina
CCBT0475	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0254	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0361	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0407	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0354	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0269	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0342	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0179	O1	Ogawa	Clinical	01/04/1992	Argentina
CCBT0145	O1	Ogawa	Clinical	13/01/1993	Argentina
CCBT0319	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0280	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0177	O1	Ogawa	Clinical	21/04/1992	Argentina

CCBT0343	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0395	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0419	O1	Ogawa	Clinical	18/02/1993	Argentina
VC016	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0244	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0314	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0219	O1	Ogawa	Clinical	19/01/1993	Argentina
CCBT0325	O1	Ogawa	Clinical	19/09/1997	Argentina
VC006	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0387	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0141	O1	Ogawa	Environmental	19/01/1993	Argentina
CCBT0130	non-O1	/	Environmental	03/06/1996	Argentina
CCBT0221	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0311	O1	Ogawa	Clinical	14/01/1993	Argentina
CCBT0298	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0453	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0149	O1	Ogawa	Clinical	01/12/1992	Argentina
CCBT0464	O1	Ogawa	Clinical	18/02/1993	Argentina
CCBT0252	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0309	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0397	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0155	non-O1	/	Environmental	11/02/2000	Argentina
CCBT0293	O1	Ogawa	Clinical	20/01/1993	Argentina
CCBT0193	O1	Ogawa	Clinical	05/05/1992	Argentina
CCBT0316	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0472	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0403	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0405	O1	Ogawa	Clinical	05/02/1993	Argentina
CCBT0406	O1	Ogawa	Clinical	12/02/1993	Argentina
CCBT0391	O1	Ogawa	Environmental	10/05/1993	Argentina
CCBT0274	O1	Ogawa	Clinical	03/02/1993	Argentina
CCBT0328	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0385	O1	Inaba	Clinical	18/02/1993	Argentina
CCBT0304	O1	Ogawa	Clinical	22/01/1993	Argentina
CCBT0206	O1	Ogawa	Clinical	16/01/1993	Argentina
CCBT0445	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0242	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0263	O1	Ogawa	Clinical	21/04/1992	Argentina
CCBT0444	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0448	O1	Ogawa	Clinical	18/02/1993	Argentina
VC013	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0362	O1	Ogawa	Clinical	02/03/1993	Argentina
CCBT0289	O1	Ogawa	Clinical	22/01/1993	Argentina

CCBT0102	non-O1	/	Environmental	12/04/2000	Argentina
CCBT0042	non-O1	/	Clinical	28/02/1996	Argentina
CCBT0079	non-O1	/	Clinical	22/04/1996	Argentina
VC034	non-O1	/	Clinical	12/05/1993	Argentina
CCBT0082	non-O1	/	Environmental	18/03/1996	Argentina
CCBT0096	/	/	Environmental	14/06/1996	Argentina
CCBT0084	O1 (genomically non-O1)	Ogawa	Clinical	26/04/1995	Argentina
CCBT0086	non-O1	/	Environmental	14/06/1996	Argentina
CCBT0081	non-O1	/	Clinical	18/04/1996	Argentina
CCBT0007	non-O1	/	Clinical	03/06/1996	Argentina
CCBT0043	non-O1	/	Clinical	28/05/1996	Argentina
CCBT0493	non-O1 non-O139	/	Clinical	15/04/2004	Argentina
CCBT0500	O1 (genomically non-O1)	Ogawa	Clinical	04/02/1993	Argentina
CCBT0090	non-O1	/	Clinical	28/05/1996	Argentina
CCBT0058	non-O1	/	Clinical	17/04/1996	Argentina
CCBT0053	non-O1	/	Clinical	06/06/1996	Argentina
VC012	non-O1	/	Environmental	17/03/1992	Argentina
CCBT0112	non-O1	/	Environmental	15/07/1992	Argentina
VC002	non-O1	/	Environmental	17/03/1992	Argentina
CCBT0106	non-O1	/	Environmental	01/12/1992	Argentina
VC001	non-O1	/	Environmental	17/03/1992	Argentina
CCBT0136	O1 (genomically non-O1)	Ogawa	Clinical	13/01/1993	Argentina
CCBT0131	O1 (genomically non-O1)	Ogawa	Environmental	19/01/1993	Argentina
CCBT0135	O1 (genomically non-O1)	Ogawa	Environmental	19/01/1993	Argentina
CCBT0063	non-O1	/	Clinical	01/09/1992	Argentina
CCBT0119	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0073	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0054	non-O1	/	Clinical	17/04/1996	Argentina
CCBT0146	non-O1	/	Environmental	29/12/1992	Argentina
VC008	non-O1	/	Clinical	28/01/1993	Argentina
CCBT0129	non-O1	/	Environmental	19/01/1993	Argentina
CCBT0162	non-O1	/	Environmental	01/12/1992	Argentina
CCBT0045	non-O1	/	Clinical	17/04/1996	Argentina
CCBT0044	non-O1	/	Clinical	17/04/1996	Argentina
CCBT0033	non-O1	/	Clinical	30/12/1996	Argentina
CCBT0060	non-O1	/	Environmental	13/11/1992	Argentina
CCBT0094	non-O1	/	Clinical	17/04/1996	Argentina
CCBT0411	O1 (genomically O1)	Ogawa	Clinical	05/02/1993	Argentina
CCBT0099	non-O1	/	Environmental	01/12/1992	Argentina
CCBT0491	/	/	Clinical	15/04/2004	Argentina
VC024	non-O1	/	Clinical	03/02/2000	Argentina

CCBT0443	non-O1	/	Clinical	02/03/1993	Argentina
CCBT0151	O1 (genomically non-O1)	Ogawa	Clinical	13/01/1993	Argentina
VC010	non-O1	/	Clinical	16/01/1993	Argentina
CCBT0087	non-O1	/	Clinical	18/04/1996	Argentina
CCBT0008	O1 (genomically non-O1)	Ogawa	Environmental	19/01/1993	Argentina
CCBT0018	non-O1	/	Clinical	03/06/1996	Argentina
CCBT0495	non-O1 non-O139	/	Clinical	18/08/2005	Argentina
CCBT0496	non-O1	/	Clinical	02/05/1994	Argentina
VC009	non-O1	/	Clinical	20/01/1993	Argentina
CCBT0093	non-O1	/	Clinical	17/04/1996	Argentina
CCBT0061	non-O1	/	Clinical	16/10/1992	Argentina
CCBT0072	non-O1	/	Clinical	19/10/1992	Argentina
CCBT0113	non-O1	/	Clinical	19/10/1992	Argentina
CCBT0494	non-O1 non-O139	/	Clinical	17/12/2003	Argentina
CCBT0101	non-O1	/	Clinical	22/12/1992	Argentina
CCBT0117	non-O1	/	Environmental	29/12/1992	Argentina
CCBT0026	non-O1	/	Clinical	08/11/1996	Argentina
CCBT0115	non-O1	/	Clinical	16/10/1992	Argentina
CCBT0020	non-O1	/	Clinical	21/01/1997	Argentina
CCBT0174	O1 (genomically O1)	Ogawa	Clinical	15/10/1997	Argentina
CCBT0118	non-O1 (genomically O1)	/	Clinical	28/05/1996	Argentina
CCBT0077	non-O1	/	Clinical	17/04/1996	Argentina
CCBT0497	O1 (genomically O1)	Ogawa	Clinical	23/02/1998	Argentina
CCBT0320	O1 (genomically O1)	Ogawa	Clinical	02/02/1993	Argentina
Mex_175	non-O1	/	Human	1993 - 1993	Guatemala
Mex_01	non-O1	/	Human	1991	Tabasco
Mex_02	non-O1	/	Human	1991	Tabasco
Mex_03	non-O1	/	Human	1991	Tabasco
Mex_04	non-O1	/	Human	1991	Tabasco
Mex_05	non-O1	/	Human	1992-93	Campeche
Mex_06	non-O1	/	Water well	1992-93	Campeche
Mex_07	non-O1	/	Water well	1992-93	Campeche
Mex_08	non-O1	/	Water well	1992	Campeche
Mex_09	non-O1	/	Human	June 1991	Hidalgo
Mex_10	non-O1	/		June1991	
Mex_11	non-O1	/	Human	July 1991	Chiapas
Mex_12	non-O1	/	Human	July 1991	Puebla
Mex_13	non-O1	/	Human	July 1991	Hidalgo
Mex_14	non-O1	/	Human	July 1991	Hidalgo
Mex_15	non-O1	/	Human	July 1991	Hidalgo
Mex_16	non-O1	/	Human	August 1991	Puebla
Mex_17	non-O1	/	Human	July 1991	Hidalgo

Mex_18	non-O1	/	Human	July 1991	Veracruz
Mex_19	non-O1	/	Human	July 1991	Hidalgo
Mex_20	non-O1	/	Environmental	August 1991	INDRE
Mex_21	non-O1	/	Human	July 1991	Veracruz
Mex_22	non-O1	/	Human	August 1991	Yucatan
Mex_23	non-O1	/	Human	August 1991	Hidalgo
Mex_24	non-O1	/	Human	August 1991	Veracruz
Mex_25	non-O1	/	Human	July 1991	Bosay
Mex_26	non-O1	/	Human	June1991	Hidalgo
Mex_27	non-O1	/	Human	August 1991	Chiapas
Mex_28	non-O1	/	Human	August 1991	Campeche
Mex_29	non-O1	/	Human	February 1992	INDRE
Mex_30	non-O1	/	Human	August 1991	Guerrero
Mex_31	non-O1	/	Human	August 1991	Chiapas
Mex_32	non-O1	/	Human	June 1991	Hidalgo
Mex_33	non-O1	/	Human	August 1991	INDRE
Mex_34	non-O1	/	Human	August 1991	INDRE
Mex_35	non-O1	/	Fish	August 1991	Veracruz
Mex_36	non-O1	/	Human	July 1991	Veracruz
Mex_37	non-O1	/	Environmental	July 1991	Veracruz
Mex_39	non-O1	/	Human	September 1991	Leon
Mex_40	non-O1	/	Human	August 1991	INDRE
Mex_38	non-O1	/	Human	August 1991	Hidalgo
Mex_41	non-O1	/	Human	September 1991	Guanajuato
Mex_42	non-O1	/	Human	August 1991	Yucatan
Mex_43	non-O1	/	Human	February 1992	Guanajuato
Mex_44	non-O1	/	Human	March 1992	Michoacan
Mex_45	non-O1	/	Human	September 1991	Chiapas
Mex_46	non-O1	/	Human	August 1991	Zacatecas
Mex_47	non-O1	/	Human	September 1991	Chiapas
Mex_48	non-O1	/	Human	September 1991	Guanajuato
Mex_49	non-O1	/	Human	August 1991	Hidalgo
Mex_50	non-O1	/	Human	August 1991	Zacatecas
Mex_51	non-O1	/	Water	1992	Tabasco
Mex_52	non-O1	/	Human	1992	Tabasco
Mex_53	non-O1	/	Human	1992	Tabasco
Mex_54	non-O1	/	Human	1992	Tabasco
Mex_55	non-O1	/	Human	1992	Tabasco
Mex_56	non-O1	/	Human	1992	Tabasco
Mex_57	non-O1	/	Human	1992	Tabasco
Mex_58	non-O1	/	Human	1992	Tabasco

Mex_59	non-O1	/	Human	1992	Tabasco
Mex_60	non-O1	/	Human	1993	Tabasco
Mex_61	non-O1	/	Human	1993	Tabasco
Mex_62	non-O1	/	Human	1992	Tabasco
Mex_63	non-O1	/	Human	1992	Tabasco
Mex_64	non-O1	/	Human	1993	Tabasco
Mex_65	non-O1	/	Human	1992	Tabasco
Mex_66	non-O1	/	Human	1992	Tabasco
Mex_67	non-O1	/	Water well	1992-93	Campeche
Mex_68	non-O1	/	Water well	1992-93	Campeche
Mex_69	non-O1	/	Water well	1992-93	Campeche
Mex_70	non-O1	/	Water well	1992-93	Campeche
Mex_71	non-O1	/	Water well	1992-93	Campeche
Mex_72	non-O1	/	Water well	1992-93	Campeche
Mex_73	non-O1	/	Water well	1992-93	Campeche
Mex_74	non-O1	/	Human	1991	Tabasco
Mex_75	non-O1	/	Human	1991	Tabasco
Mex_76	non-O1	/	Human	1991	Tabasco
Mex_77	non-O1	/	Human	1991	Tabasco
Mex_78	non-O1	/	Human	1991	Tabasco
Mex_79	non-O1	/	Human	1991	Tabasco
Mex_80	non-O1	/			
Mex_81	non-O1	/			
Mex_82	non-O1	/			
Mex_83	non-O1	/			Campeche
Mex_84	non-O1	/		apr-oct 1993	
Mex_85	non-O1	/		apr-oct 1993	
Mex_88	non-O1	/	Food	1993	Hermosillo
Mex_89	non-O1	/	Water	1993	Hermosillo
Mex_90	non-O1	/	Food	1993	Hermosillo
Mex_91	non-O1	/	Fish	1993	Hermosillo
Mex_92	non-O1	/	Fish	1993	Hermosillo
Mex_93	non-O1	/	Fish	1993	Hermosillo
Mex_94	non-O1	/	Water	1993	Hermosillo
Mex_95	non-O1	/	Fish	1993	Hermosillo
Mex_96	non-O1	/	Water	1993	Hermosillo
Mex_97	non-O1	/	Fish	1993	Hermosillo
Mex_98	non-O1	/	Fish	1993	Hermosillo
Mex_99	non-O1	/	Fish	1993	Hermosillo
Mex_100	non-O1	/	Fish	1993	Hermosillo
Mex_101	non-O1	/	Shrimp	1993	Hermosillo
Mex_102	non-O1	/	Fish	1993	Hermosillo
Mex_103	non-O1	/	Fish	1993	Hermosillo

Mex_104	non-O1	/	Fish	1993	Hermosillo
Mex_105	non-O1	/	Fish	1993	Hermosillo
Mex_106	non-O1	/	Fish	1993	Hermosillo
Mex_107	non-O1	/	Sewage	1993	Hermosillo
Mex_108	non-O1	/	Sewage	1993	Hermosillo
Mex_109	non-O1	/	Sewage	1993	Hermosillo
Mex_110	non-O1	/	Septic tank	1993	Hermosillo
Mex_111	non-O1	/	Sewage	1993	Hermosillo
Mex_179	non-O1	/	Sewage	1993	Hermosillo
Mex_178	non-O1	/	Sewage	1993	Hermosillo
Mex_112	non-O1	/	Sewage	1993	Hermosillo
Mex_113	non-O1	/	Sewage	1993	Hermosillo
Mex_115	non-O1	/	Dam	1993	Hermosillo
Mex_114	non-O1	/	Dam	1993	Hermosillo
Mex_116	non-O1	/	Dam	1993	Hermosillo
Mex_117	non-O1	/	Sewage	1993	Hermosillo
Mex_118	non-O1	/	Sewage	1993	Hermosillo
Mex_119	non-O1	/	Sewage	1993	Hermosillo
Mex_120	non-O1	/	Sewage	1993	Hermosillo
Mex_121	non-O1	/	Sewage	1993	Hermosillo
Mex_176	non-O1	/	Sewage	1993	Hermosillo
Mex_122	non-O1	/	Sewage	1993	Hermosillo
Mex_123	non-O1	/	Sewage	1993	Hermosillo
Mex_124	non-O1	/	Sewage	1993	Hermosillo
Mex_125	non-O1	/	Sewage	1993	Hermosillo
Mex_177	non-O1	/	Sewage	1993	Hermosillo
Mex_126	non-O1	/	Sewage	1993	Hermosillo
Mex_127	non-O1	/		apr-oct 1993	Mexico
Mex_128	non-O1	/			Mexico
Mex_129	non-O1	/			Mexico
Mex_130	non-O1	/			Mexico
Mex_131	non-O1	/			Mexico
Mex_132	non-O1	/			Mexico
Mex_133	non-O1	/			Mexico
Mex_134	non-O1	/			Mexico
Mex_135	non-O1	/			Mexico
Mex_136	non-O1	/			Mexico
Mex_137	non-O1	/			Mexico
Mex_138	non-O1	/			Mexico
Mex_139	non-O1	/			Mexico
Mex_141	non-O1	/			Mexico
Mex_142	non-O1	/			Mexico
Mex_143	non-O1	/			Mexico

Mex_144	non-O1	/			Mexico
Mex_145	non-O1	/	Human	1992 - 1993	Guatemala
Mex_146	non-O1	/	Human	1992 - 1993	Guatemala
Mex_147	non-O1	/	Human	1992 - 1993	Guatemala
Mex_148	non-O1	/	Human	1992 -1993	Guatemala
Mex_174	non-O1	/	Human	1992 - 1993	Guatemala
Mex_149	non-O1	/	Human	1992 - 1993	Guatemala
Mex_173	non-O1	/	Human	1992 - 1993	Guatemala
Mex_151	non-O1	/	Human	1992 - 1993	Guatemala
Mex_152	non-O1	/	Human	1992 - 1993	Guatemala
Mex_172	non-O1	/	Human	1992 - 1993	Guatemala
Mex_153	non-O1	/	Human	1992 - 1993	Guatemala
Mex_171	non-O1	/	Human	1992 - 1993	Guatemala
Mex_154	non-O1	/	Human	1992 - 1993	Guatemala
Mex_170	non-O1	/	Human	1992 - 1993	Guatemala
Mex_155	non-O1	/	Human	1992 - 1993	Guatemala
Mex_169	non-O1	/	Human	1992 - 1993	Guatemala
Mex_156	non-O1	/	Human	1992 - 1993	Guatemala
Mex_168	non-O1	/	Human	1992 - 1993	Guatemala
Mex_157	non-O1	/	Human	1992 - 1993	Guatemala
Mex_158	non-O1	/	Human	1992 - 1993	Guatemala
Mex_167	non-O1	/	Human	1992 - 1993	Guatemala
Mex_159	non-O1	/	Human	1992 - 1993	Guatemala
Mex_160	non-O1	/	Human	1992 - 1993	Guatemala
Mex_161	non-O1	/	Human	1992 - 1993	Guatemala
Mex_162	non-O1	/	Human	1992 - 1993	Guatemala
Mex_163	non-O1	/	Human	1992 - 1993	Guatemala
Mex_164	non-O1	/	Human	1992 - 1993	Guatemala
Mex_165	non-O1	/	Human	1992 - 1993	Guatemala
Mex_86	non-O1	/			
Mex_87	non-O1	/			
NCTC_30	Non-O1	/	/	1916	See table 5.2
NCTC_3661	O1	/	/	1931	See table 5.2
NCTC_4693	O1	/	/	Pre-1936	See table 5.2
NCTC_4711	Non-O1/O139	/	/	Pre-1936	See table 5.2
NCTC_4714	Non-O1/O139	/	/	1934	See table 5.2
NCTC_4715	Non-O1/O139	/	/	1934	See table 5.2

NCTC_4716	Non-O1/O139	/	/	1932	See table 5.2
NCTC_5395	O1	/	/	1938	See table 5.2
NCTC_5596	O1	/	/	Pre-1939	See table 5.2
NCTC_6585	O1	/	/	Pre-1944	See table 5.2
NCTC_7258	O1	/	/	Pre-1948	See table 5.2
NCTC_7260	O1	/	/	Pre-1948	See table 5.2
NCTC_7270	O1	/	/	Pre-1948	See table 5.2
NCTC_8021	O1	/	/	Pre-1950	See table 5.2
NCTC_8039	O1	/	/	Pre-1950	See table 5.2
NCTC_8040	O1	/	/	Pre-1950	See table 5.2
NCTC_8041	O1	/	/	Pre-1950	See table 5.2
NCTC_8042	Non-O1/O139	/	/	Pre-1950	See table 5.2
NCTC_8050	O1	/	/	Pre-1950	See table 5.2
NCTC_8367	O1	/	/	Pre-1952	See table 5.2
NCTC_8457	O1	/	/	1930	See table 5.2
NCTC_9420	O1	/	/	Pre-1955	See table 5.2
NCTC_9421	O1	/	/	Pre-1955	See table 5.2
NCTC_9422	O1	/	/	Pre-1955	See table 5.2
NCTC_9423	O1	/	/	Pre-1955	See table 5.2
NCTC_10255	O1	/	/	1961	See table 5.2
NCTC_10256	O1	/	/	1961	See table 5.2
NCTC_10732	O1	/	/	1952	See table 5.2
NCTC_10733	O1	/	/	1952	See table 5.2
NCTC_10954	O1	/	/	1973	See table 5.2
NCTC_11090	O1	/	/	Pre-1950	See table 5.2
NCTC_11348	Non-O1/O139	/	/	Pre-1981	See table 5.2
NCTC_11500	Non-O1/O139	/	/	Pre-1983	See table 5.2

NCTC_11643	Non-O1/O139	/	/	Pre-1985	See table 5.2
NCTC_12946	O139	/	/	1993	See table 5.2