1. Introduction

1.1 Cholera

1.1.1 Overview

Cholera is regarded by many as a disease of great historical importance that marks many of the leaps we have made in the understanding of infectious diseases. Like many diseases of historical significance such as plague, the spread of cholera was also believed to be *via* bad air or ‘miasma’. It was not until the reports of John Snow’s findings in 1854 that a connection between contaminated drinking water and cholera began to be recognised. John Snow showed that most cholera deaths in a particular region of London were clustered around an area where people acquired water from the same pump on Broad Street. He showed that if an individual cholera victim lived away from the Broad Street vicinity, they did sample this specific pump because they sometimes preferred the taste of the water from there. The removal of the handle of Broad Street pump and the resultant drop in the cases of cholera is heralded by many as the beginning and birthplace of the field of Epidemiology. The identification of *Vibrio cholerae* as the etiological agent of cholera was made by Robert Koch in 1894, soon after he proposed the ‘germ theory of infection’ or as it is called today, ‘Koch’s Postulate’.

Though clearly an ancient disease, cholera is still common in many regions of the world despite having one of the simplest known treatment regimes: oral rehydration. Moreover, since 2007, the incidence of cholera has gradually increased and the World Health Organisation (WHO) reported 317534 cases and 7543 deaths in 2010 (WHO weekly epidemiological records 2008 and 2011). Since current WHO guidelines no longer require notification of cholera cases, the true burden of the disease is only estimated but is believed to be in the millions every year. For example, not even a single case of cholera has been reported from Bangladesh since 2009, a situation clearly far from the true incidence level.

In an attempt to acknowledge the dire global situation relating to cholera, a new resolution has recently been adopted by the WHO for an integrated and comprehensive global approach to the disease (http://www.who.int/cholera/technical/Resolution_CholeraA64_R15-en.pdf). The concern is valid because cholera is a toxin-mediated disease that involves rapid onset of severe watery diarrhea that can lead to the death of a patient within hours if rehydration therapy is not promptly administered.

The historical and current impact of cholera on humanity and in generally shaping societies, especially in the developing world, is arguably enormous and this is evident from the mention of the disease in old literature and novels written around it (Sack, *et al.*, 2004). “Asiatic cholera”, as it was once called, has now spread globally in the form of epidemics or pandemics and even today it is on the verge of becoming endemic in several countries that generally face hygiene and sanitation problems or are suffering the aftermath of natural disasters, such as Haiti (2010; Barzilay, *et al.*, 2013; Chin, *et al.*, 2011). Several vaccines for protection against cholera are available in the international market but they are not extensively used in low income countries (Lopez, *et al.*, 2008). Therefore the current best approach to cholera control is improvement of hygiene and sanitation where it is most needed, alongside active monitoring of outbreaks in both endemic and epidemic settings. Currently we lack a detailed understanding as to how *V. cholerae* moves around and evolves, although water is clearly a critical factor. In areas where complete and sustained access to clean water is missing, a better understanding of the bacterium and its’ general epidemiology is required. Recently, partly provoked by a high profile cholera outbreak in Haiti, as well as ongoing efforts, scientists around the globe have become aware of the paramount importance of continuous and retrospective surveillance using accurate systems such as molecular technologies for tracking and understanding this disease.

1.1.2 Cholera Pandemics

Since the existence of records, cholera has been endemic in South Asia, mainly in the regions bordering the Bay of Bengal in India and Bangladesh. A classical example of how much cholera was feared by the population of Kolkata is that a cholera temple was built as a refuge to protect people from the disease (Sack, *et al.*, 2004). The public health system and general situation has improved but the region is still considered the reservoir for outbreaks of *V. cholerae*. Historically, these two countries have accounted for significantly higher mortality rates of cholera compared to the other regions of the world.

There have been seven acknowledged pandemics of cholera to date and the world is currently still experiencing the seventh (Figure 1.1). It is believed that the first of the seven reported cholera pandemics started in 1817 and spread from the Indian sub-continent to Russia along trade routes (Sack, *et al.*, 2004).

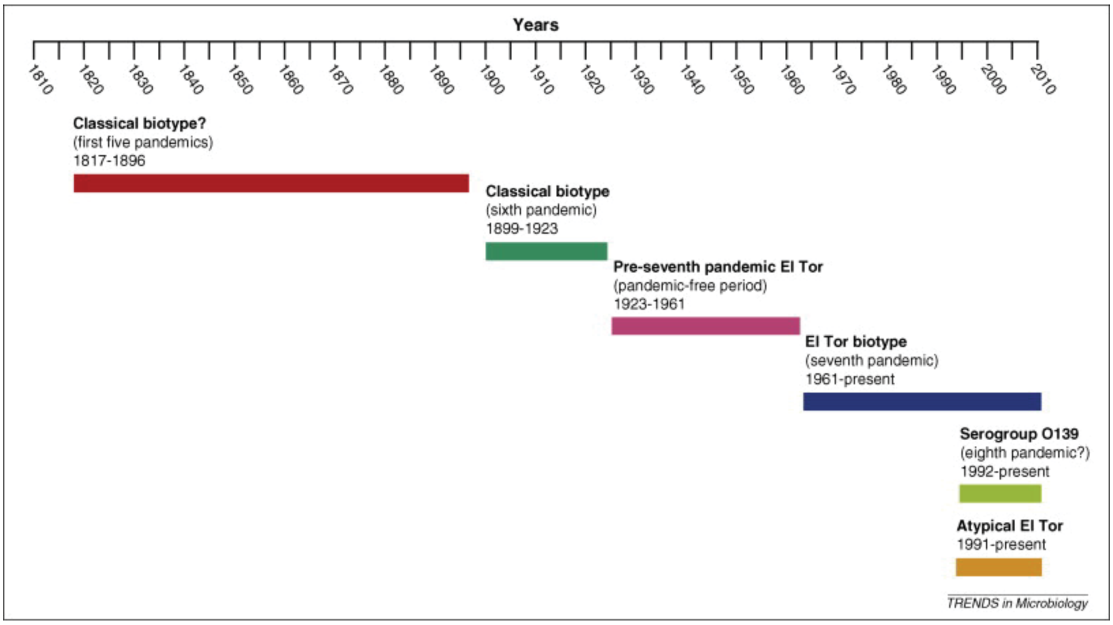

Figure 1.1: Timeline showing all the seven pandemics and hypothesized eighth pandemic or seventh sub-pandemic. Figure reproduced from (Safa, *et al.*, 2010).

The second pandemic started in 1826 and reached the major European cities in 1830 and London in 1831. This pandemic was also traced back to the Bay of Bengal region when the parallels between the movement of workers who were brought into Great Britain for coal mining during the industrial revolution and the appearance of cases were linked. The next three pandemics affected almost the whole world including countries in Asia, Africa, Americas, Europe and even Australia (Sack, *et al.*, 2004). The fifth pandemic ended in 1896 and it was only in 1894 that association of *V. cholerae* as the causative agent of this severe diarrheal disease was established. The sixth pandemic started in 1899, began receding in 1923 and some believe that it lasted up to the better part of 1925. Since the agent responsible for the disease was known by this time, *V. cholerae* collections from this period do exist in the historical archives of several research organizations around the world. The isolates of that period were predominantly typed as so called "classical" *V. cholerae* (see section 1.2.1). A map showing actual and supposed routes of spread of so called 'Asiatic cholera' for the first five pandemics was produced in 1885 by John C Peters of the United States. The map shown in Figure 1.2 illustrates the spread of cholera from Hindoostan (the then name for the Indian sub-continent) to the rest of the world. This map has strong resonance with the work outlined in this thesis, however, even now the mechanism by which cholera has spread across the world has been controversial. Despite several maps like the one illustrated, some people still believe that cholera evolves locally and independently of a source in the Bay of Bengal.

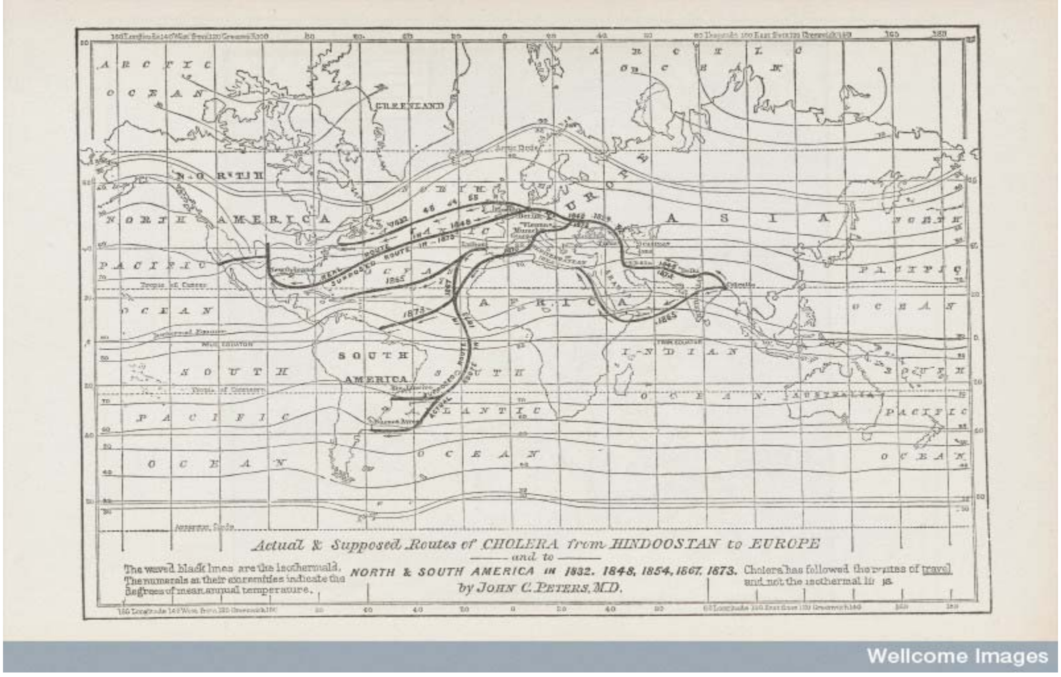
A R C T I C
O C E A N
GREENLAND
N O R T H
A M E R I C A
E U R O P E
A S I A
A F R I C A
P A C I F I C
O C E A N
S O U T H
A M E R I C A
S O U T H
A T L A N T I C
O C E A N
I N D I A N
O C E A N
A U S T R A L I A
P A C I F I C
O C E A N
Actual & Supposed Routes of CHOLERA from HINDOOSTAN to EUROPE
and to
NORTH & SOUTH AMERICA in 1832, 1848, 1854, 1867, 1873.
by JOHN C. PETERS, M.D.
The waved black lines are the isothermals. The numerals at their extremities indicate the degrees of mean annual temperature.
Cholera has followed the routes of travel and not the isothermal lines.
Wellcome Images

Figure 1.2: One of the first maps showing the spread of Asiatic cholera from the South Asian sub-continent to the rest of the world (Wellcome Image Library reference GC WC262 1885W47t).

Between the cholera outbreaks of sixth and seventh pandemic, there was a latent period (1923-1961) when the classical strains were not causing cholera outbreaks at any significant level. This latency was broken by an outbreak in 1961, which was caused by strains showing biochemical traits different from those of the classical biotype (Safa, *et al.*, 2010). It was noted that the phenotypic, biochemical and microbiological features of these new strains matched those of *V. cholerae* isolated from pilgrims in a village in Egypt known as El Tor, who were travelling to Mecca in 1905 (Sack, *et al.,* 2004). From 1961 onwards, the spread of *V. cholerae* of this biotype was significantly quicker than the previous pandemics and by 1991 it was reported from almost all parts of the world, including a severe outbreak over much of Latin America, a continent that had not experienced the calamity of previous pandemics at such scale.

1.2 *Vibrio* bacteria

Vibrio is the name given to the genus of bacteria that fall into the family Vibrionaceae. This name is derived from Filippo Paccini's work in 1854 when he isolated coma-shaped motile microorganisms and called them 'vibrions'. Vibrionaceae consists of three genera *Vibrio, Photobacterium and Salinivibrio,* but *Vibrio* is the type genus of this family. Vibrios are Gram-negative straight or slightly curved motile rods, which have flagella and are mainly found in aquatic reservoirs such as fresh or brackish water, estuaries, rivers and coastal waters. In these habitats they have been shown to be associated with copepods, algae, zooplankton, phytoplankton and are also found on the surface of shellfish. Bacteria of *Vibrio* genus are also known to form biofilms on the surface of crustaceans where it is thought they can better resist the environmental stress and natural antibiotics while maintaining nutrient absorbtion. Members of some species of the *Vibrio* genus (*V. vulnificus, V. harveyi, V. parahaemolyticus, V. anguillarum* and *V. cholerae*) can be bioluminescent and live in mutualistic association with fish, frog and other marine life forms. This life-style can aid chemical or photo communication, a phenomenon called 'quorum sensing', between bacteria and animals alike. Other Vibrios have pathogenic potential, causing mainly enteric diseases and sometimes infection of open wounds and even septicaemia. *Vibrio cholerae*, *Vibrio mimicus*, *Vibrio parahaemolyticus* and *Vibrio vulnificus* are four species in *Vibrio* genus that are known to cause clinically significant disease in humans. They can be distinguished from enteric pathogens of family *Enterobacteriaceae* by an oxidase test since vibrios are oxidase positive, have O_2 as a universal electron acceptor and they test negative for denitrification. Even those species that cause clinically similar disease can exploit different pathogenic mechanisms and pathways; for example *V. parahaemolyticus* is potentially invasive and affects colon whereas *V. cholerae* releases a powerful enterotoxin in the small intestine, which can stimulate severe diarrhoea.

According to Bergey's Manual of Systematic Bacteriology (Gammaproteobacteria, 2nd edition 2B, 2005) most vibrios are facultative anaerobes and can grow in synthetic media with glucose as the source of energy and carbon. Most Vibrio species require slightly alkaline (2-3% NaCl or sea water) conditions (*V. cholerae* and *V. mimicus*) but there are some that are halophiles (*V. parahaemolyticus* and *V. vulnificus*). A few species grow at temperatures below 25 °C but most grow well between 25-37 °C. The molecular GC content of their DNA ranges between 38-51%.

1.2.1 *V. cholerae*: the species and classification

Domain: Bacteria

Phylum: Proteobacteria

Class: Gammaproteobacteria

Order: Vibrionales

Family: Vibrionaceae

Genus: Vibrio

Species: *Vibrio cholerae*

Serogroups: Over 200

Epidemic serogroups: O1 and O139

Pandemic serogroup: O1

O1 serogroup biotypes: classical and El Tor

O1 serogroup serotypes: Inaba and Ogawa

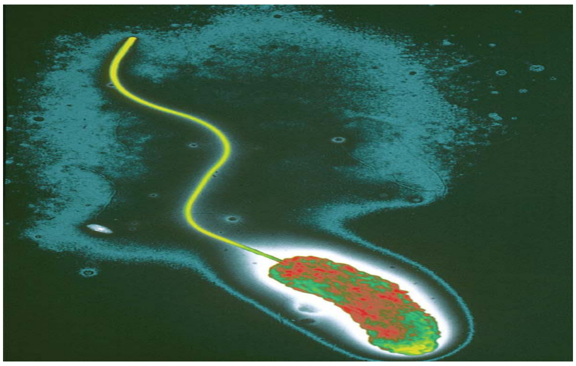

Figure 1.3: 10,000 x magnification of *V. cholerae* bacterium. Figure reproduced from Waldor *et al.* 2000 (Waldor and RayChaudhuri, 2000).

V. cholerae is the type species of the genus *Vibrio*. Since this species causes cholera, historically one of the most important diseases alongside plague and typhoid, it is the most extensively studied in the family Vibrionaceae. *V. cholerae* requires slightly alkaline conditions for optimum growth and can grow in conditions up to a maximum of pH 10 but the bacteria does not grow well below pH 6. The bacterial cell is comma shaped and has sheathed polar flagellum that it uses for motility as shown in Figure 1.3.

Like many bacterial pathogens, individual *V. cholerae* can be distinguished by the antigenic composition of their lipopolysaccharides (LPS). While the lipid-A and core-PS have generally similar structures within the species and are responsible for endotoxicity in different serogroups of *V. cholerae*, the O-antigen polysaccharide can have distinct structures and the molecule is involved in immunogenicity and induction of vibriocidal antibodies in the mammalian host (Chatterjee and Chaudhuri, 2004; Chatterjee and Chaudhuri, 2006). The O-antigen is the outermost region of the LPS on the surface of the Vibrio and the epitopes associated with this antigen class have been used to sub-divide the O1 serogroup bacteria into Inaba, Ogawa and Hikojima serotypes using specific typing sera.

It is important to note that the epidemics and pandemics of cholera have been caused by *V. cholerae* of either the O1 or O139 serogroup, despite the fact that based on the O-antigen variation more than 200 O serogroups have been identified. O1 antigenic forms of *V. cholerae* have likely predominated throughout all seven pandemics.

Strains of serogroup O139 raised a concern in 1992 when they emerged as a novel clade, when the number of cholera cases due to this serogroup surpassed the O1 cases in Bangladesh (Faruque, *et al.*, 2003). However, since then, O139-positive isolates have ceased to compete with the O1 strains, in terms of disease causation, and have largely disappeared. O139 isolates harbor a genome related to typical O1 El Tor *V. cholerae* except that the gene cluster encoding their O-antigen has been replaced with a different set of genes from normal El Tor isolates (Mutreja, *et al.*, 2011).

Most O1 and O139-positive *V. cholerae* clinical isolates can produce a potent enterotoxin called cholera toxin (CT; discussed in detail in section 1.2.8) that is responsible for the signature rice water stool during episodes of cholera disease. The general surface antigen features of *V. cholerae* are nicely illustrated in Figure 1.4.

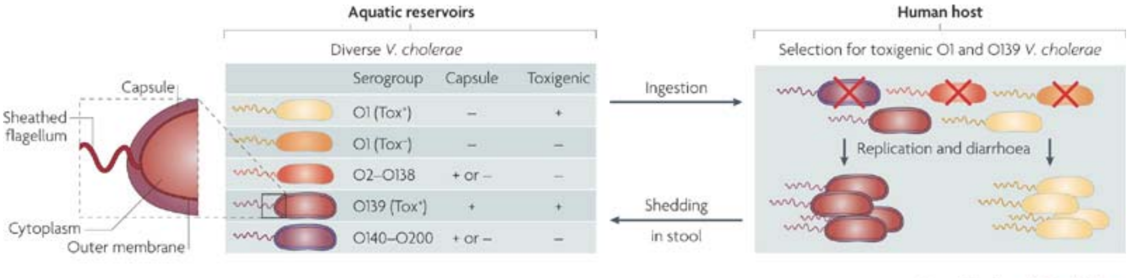

Figure 1.4: O-antigen serogroups, their main properties and the selection they may undergo when they infect humans are shown. Figure reproduced from Nelson *et al.* 2009 (Nelson, *et al.*, 2009).

V. cholerae can be further sub-divided into two biotypes known as classical and El Tor based on several phenotypic, biochemical, molecular and genomic differences (listed in Table 1).

Biotype	**Biochemical**					**Genotypic**		
	CCA	**PB**	**VP**	**Phage IV**	**Phage 5**	***tcpA***	***rstR***	***ctxB***
Classical	-	s	-	s	r	cla	cla	Type 1
El Tor	+	r	+	r	s	ET	ET	Type 3

Table 1: Table (sourced from (Safa, et al., 2010)) showing biochemical and genotypic

properties traditionally used to differentiate between classical and El Tor biotypes of O1 *V. cholerae*. CCA – chicken cell agglutination; PB – Polymyxin B test; VP – Voges-Proskauer test; s – sensitive; r – resistant; cla – classical; ET – El Tor. tcpA, rstR and ctxB represent the genes encoding toxin co-regulated pilin, transcriptional regulator of CT and sub-unit B of CT, respectively

Depending on the techniques available, *V. cholerae* isolates can be subtyped further based on either their cholera toxin sequence or differential methylation patterns on their LPS O-antigen, or a combination of the two. However, many different variants have now been found that do not fit under any of these categories (Chun, *et al.*, 2009). The details of this classification are discussed in section 1.2.8.

1.2.2 Ecology of *V. cholerae*

It is now clear that cholera does not just spread *via* fecal oral transfer between humans in food and drinking water from infected to uninfected individuals (Sack, *et al.*, 2004). Evidence (Faruque, *et al.*, 2005) has emerged from cholera-endemic areas that O1 and O139 *V. cholerae* can live in brackish water possibly on the surface of zooplanktons and phytoplanktons in mutual association. Taking into account these complex host-pathogen-environmental interactions, the whole life cycle of *V. cholerae* becomes a vital element in understanding the disease (Figure 1.5).

Thus, there are arguably two phases of the *V. cholerae* life cycle; one in the aquatic environment where they exist as free swimming cells in brackish water or in association with crustaceans, green algae, copepods or on the surfaces of shell fish and crabs (Colwell, 1996; Islam, *et al.*, 1994). The second phase is inside the human host at the luminal surface of the small intestine, where they multiply and release toxin.

On some surfaces *V. cholerae* can form biofilms that may facilitate their persistence in water currents, potentially maintaining a supply of nutrients between human cholera epidemics (Watnick, *et al.*, 2001). It has been proposed that environmental *V. cholerae* can exist in a viable but non-culturable (VBNC) phase i.e. they stay metabolically active and respire but cannot be readily cultured (Colwell, 2000).

Interactions between bacteriophages and *V. cholerae* in the aquatic systems have also been proposed to be important in containing the numbers of *V. cholerae* in these habitats (Faruque, *et al.*, 2005).

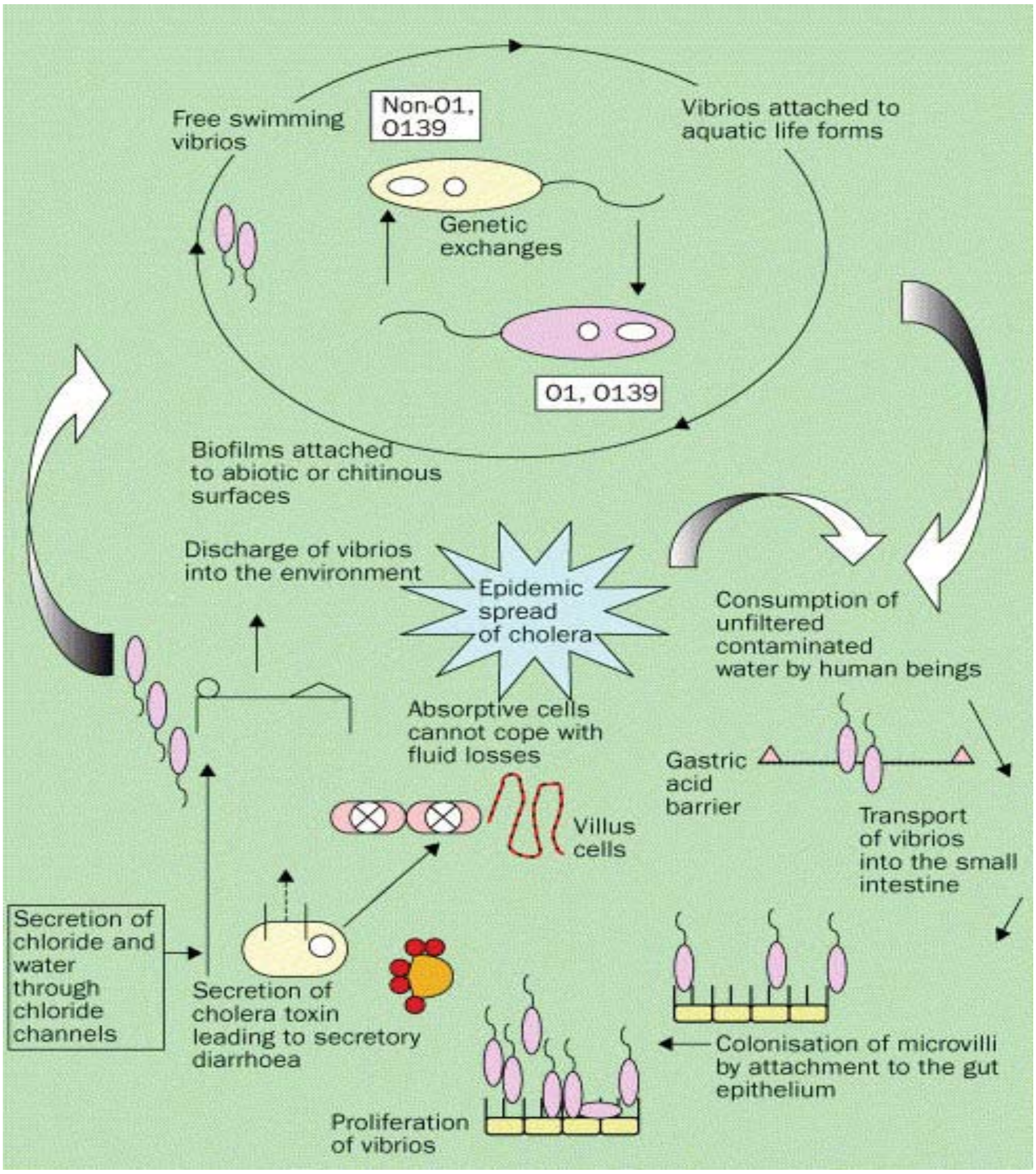
Free swimming vibrios
Non-O1, O139
Vibrios attached to aquatic life forms
Genetic exchanges
O1, O139
Biofilms attached to abiotic or chitinous surfaces
Discharge of vibrios into the environment
Epidemic spread of cholera
Consumption of unfiltered contaminated water by human beings
Absorptive cells cannot cope with fluid losses
Gastric acid barrier
Villus cells
Transport of vibrios into the small intestine
Secretion of chloride and water through chloride channels
Secretion of cholera toxin leading to secretory diarrhoea
Colonisation of microvilli by attachment to the gut epithelium
Proliferation of vibrios

Figure 1.5: The proposed life cycle of *V. cholerae*. Figure reproduced from Sack *et al.* 2004 (Sack, *et al.*, 2004).

It has been shown that an increase in numbers of cholera cases, can be associated with a subsequent rise in bacteriophages that can be isolated from the environment (Faruque, *et al.*, 2005). This delayed concordance has been said to be important in containing the concentration of the outbreak associated *V. cholerae*, thereby eventually causing a decline in the outbreak. Dual-peak cholera outbreaks in metropolitan cities of Bangladesh are well described and a study at the International Centre for Diarrheal Disease Research (ICDDRB) showed that the number of lytic bacteriophages isolated from the stools of patients increased with the rise in number of patients reporting to the hospital (Faruque, *et al.*, 2005). Moreover, 10^2 to 10^8 bacteriophages have been titred in rice water stools during peak cholera seasons. However, it is not yet understood why such high concentration of lytic bacteriophages is unable to totally clear *V. cholerae* infections from human gut. Some scientists have proposed that this failure may be important for the increased propagation and clonal expansion of bacteriophages during outbreaks (Faruque, *et al.*, 2005).

1.2.3 Epidemiology

Cholera is a disease of poverty and lack of hygiene. The spread of cholera is generally associated with contaminated drinking water and food (primary), but the transmission of the disease can be on going through the infected individuals (secondary). While the former or the primary transmission is common in endemic areas, the later or the secondary mode of cholera transmission can seed epidemics in any non-endemic region.

The *V. cholerae* inoculum size needed for typical cholera disease is regarded as very high (10^8 in healthy people) but as low as 10^5 bacteria are sufficient for causing disease in malnourished individuals with low gastric acid production capability (Hornick, *et al.*, 1971; Sack, *et al.*, 1998). It is believed that during outbreak situations, the infectious dose is even lower because the *V. cholerae* strains can take on a hyper-virulent phenotype after several passages through the human gut-environment cycle (Butler, *et al.*, 2006; Larocque, *et al.*, 2005; Merrell, *et al.*, 2002). The size of pathogenic inoculum in the aquatic reservoirs may also depend upon the season (Pascual, *et al.*, 2000). For instance, in the Indian sub-continent, cholera peaks during warm periods before, during and after the monsoon rain falls. In Bangladesh, cholera generally peaks twice in a year in Dhaka and one annual peak is seen in Northern and Eastern Bangladesh (Sack, *et al.*, 2004). In Latin America too, El Nino events (driven by warm ocean currents) have been linked to the cycles of cholera outbreaks (Mandal, *et al.*, 2011). The model of the interactions of *V. cholerae* with human and environmental hosts is shown in Figure 1.6.

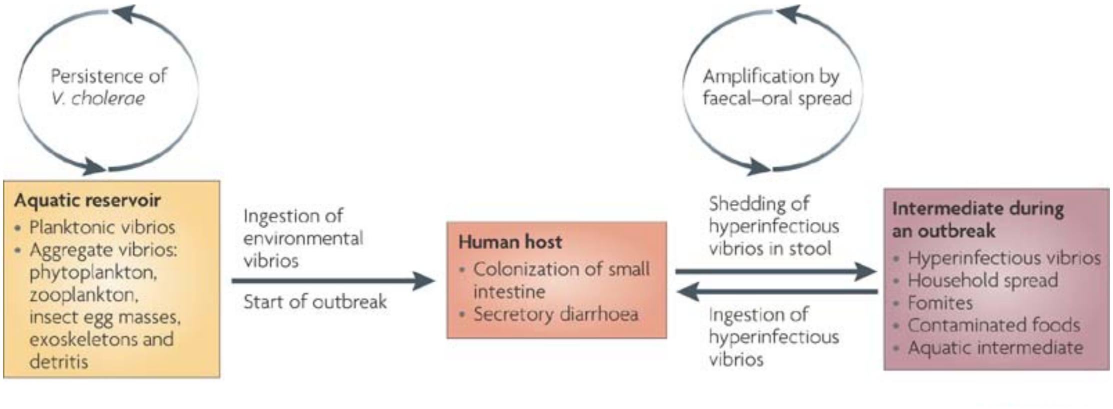
Persistence of V. cholerae
Amplification by faecal–oral spread
Aquatic reservoir
• Planktonic vibrios
• Aggregate vibrios: phytoplankton, zooplankton, insect egg masses, exoskeletons and detritis
Ingestion of environmental vibrios
Start of outbreak
Human host
• Colonization of small intestine
• Secretory diarrhoea
Shedding of hyperinfectious vibrios in stool
Ingestion of hyperinfectious vibrios
Intermediate during an outbreak
• Hyperinfectious vibrios
• Household spread
• Fomites
• Contaminated foods
• Aquatic intermediate

Figure 1.6: Circulation of pathogenic *V. cholerae* between environment and humans during outbreaks. Figure reproduced from Nelson *et al.* 2009 (Nelson, *et al.*, 2009).

Host susceptibility has also been linked to the incidence of the disease. It has been shown that people with O blood groups are at much higher risk of getting severe cholera following El Tor *V. cholerae* infection as compared to other blood groups. A study of population in Bangladesh found that there are less than the statistically predicted number of people with O blood group, which may be due to natural selection against the allele in that region (Glass, *et al.*, 1985; Harris, *et al.*, 2008). Another study has proposed that people with a particular long, palate, lung and nasal epithelium carcinoma associated protein 1 (LPLUNC 1) variant expressed in the epithelial cells of small intestine show severe cholera disease due to compromised innate immune responses initiation against *V. cholerae* (Flach, *et al.*, 2007). Lack of micronutrients like vitamin A and zinc has also been shown to have positive correlation with the ease with which *V. cholerae* infection could take place during an outbreak (Roy, *et al.*, 2008). Hence, oral zinc is given to children with severe diarrhea to reduce the stool volume and control diarrhea. Undoubtedly, acquired immunity can also contribute to resistance to cholera. Individuals become more resistant to cholera as they grow older and vibriocidal antibodies have been associated with this immunity (Glass, *et al.*, 1982). A link between age and the population affected by cholera has been reported as in an endemic area, the worst affected age groups are 2 to 4 year old children and old age people whereas in an unexposed population all individuals have equal chances of getting the infection (Glass, *et al.*, 1982). Since cholera outbreaks can become epidemics covering regions where trade and travel occur, cases that meet the clinical definition of cholera should be accurately reported to the relevant departments so that prompt public health action can be taken and the population can be forewarned.

1.2.4 *V. cholerae* infection and symptoms

Cholera is a predominantly non-invasive disease of the small intestine. For *V. cholerae* infection to take place, the bacteria ingested with the contaminated food or water inoculum must succeed in crossing the gastric acid barrier in the stomach before being able to colonize the small intestine. Epidemic O1 *V. cholerae* can express toxin co-regulated pili (TCP) at their surfaces, which interact with the receptors on the mucosal cells in the intestine and serve a paramount role in colonization (Faruque and Mekalanos, 2003; Manning, 1997). Bacteria penetrate the mucus layer overtop the mucosa by utilizing their flagellar motility machinery and once the adherence is complete, cholera enterotoxin (CT) can be delivered to the mucosal cells efficiently, initiating water loss and in turn the typical cholera disease.

The symptoms (Figure 1.7) of the disease normally appear after an incubation period ranging between 18 hours and 5 days (Sack, *et al.*, 2004). At first, the patient may develop mild watery diarrhea, which can be quickly followed by vomiting. In severe cases, abruptly, huge volumes of stool that resembles rice-water are discharged involuntarily. The rate of dehydration can be so severe that in severe cases the fluid loss may reach 0.5L to 1L per hour.

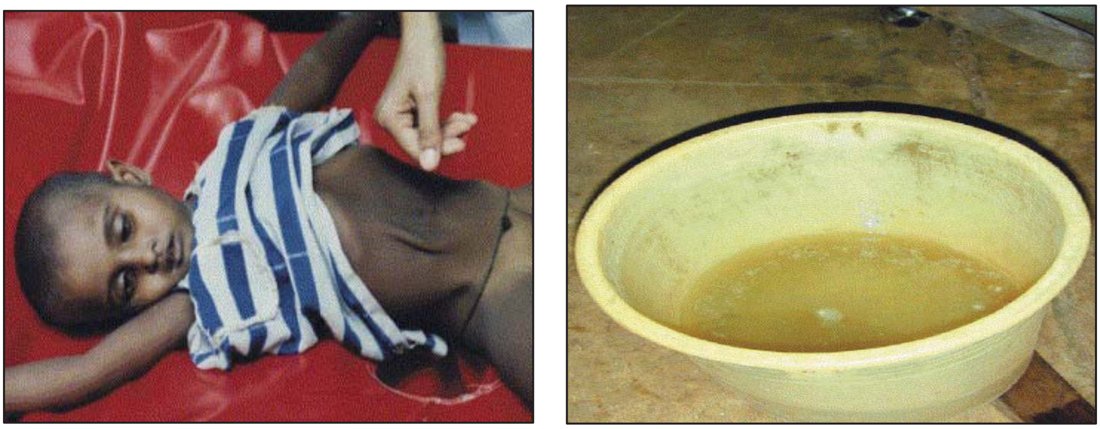

Figure 1.7: A suffering child shows typical cholera symptoms; rice water stool is collected for diagnosis and monitoring of diarrheal volume discharged. Figures reproduced from Sack *et al.* 2004 (Sack, *et al.*, 2004).

This rapid loss of body fluids can lead to a worryingly low blood pressure, low peripheral pulse, muscle cramps, decreased skin turgor, sinking of eyes and wrinkled limbs. If the fluid loss is not compensated immediately, death may occur in a matter of a few hours. However, an immediate treatment based on accurate diagnosis and monitoring can revive the patient surprisingly quickly. In endemic areas, electrolyte imbalance and hypoglycaemia is common in patients but can be corrected with intravenous administration of saline fluids (Mandal, *et al.*, 2011).

1.2.5 Diagnosis

As soon as the above cholera symptoms discussed in section 1.2.4 are noticed, cholera treatment centers should not wait for the diagnoses results to confirm the presence of *V. cholerae* and retrospective rehydration therapies should commence. The fecal samples for cholera confirmation should preferably be sent to diagnostic and reference labs in Cary-Blair transport medium to avoid any bacterial survival loss (Sack, *et al.*, 2004).

Rapid diagnosis can be performed under dark field microscope where stool samples are investigated for the presence of ‘darting’ microorganisms that freeze on addition of O1 or O139 antiserum. Rapid immunoassays are also commercially available and are mainly used in monitoring the spread of an ongoing outbreak. PCR and DNA probe tests have also been developed to detect existing and known variant strain types, mainly in non-endemic areas to establish the genealogy of the outbreak but occasionally in endemic settings too.

In parallel, the fecal specimen can be inoculated into alkaline peptone water and plated onto thiosulphate citrate bile salts sucrose (TCBS) agar. While TCBS is selective for *V. cholerae* and restricts the growth of other microbes on the plate, alkaline peptone water is an enrichment broth that promotes the growth of *V. cholerae* on overnight incubation. Alkaline peptone water is especially useful when patients report with mild diarrhea or when an environmental sample is being examined for *V. cholerae*. TCBS plates are incubated for 24 hours and *V. cholerae* appear as very distinctive yellow, raised-centre colonies as seen in Figure 1.8.

Typical *V. cholerae* colonies can be further tested biochemically for oxidase positivity and denitrification. To distinguish between the O1 and O139 serogroups, agglutination with specific antisera is carried out. However, this must be done on colonies taken from non-selective media since TCBS colonies can give false positive results. All culture positive specimens that agglutinate with either O1 or O139 must be reported to the relevant reference laboratory and health departments for true estimation of the scale of the outbreak.

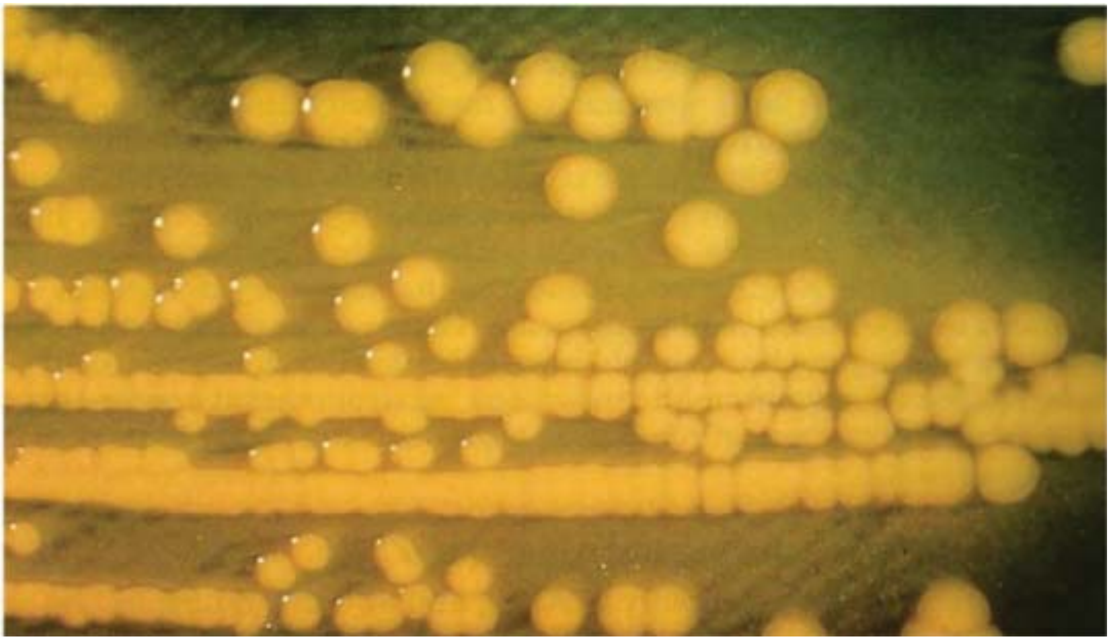

Figure 1.8: *V. cholerae* colonies on selective TCBS media. Figure sourced from Laboratory methods for the Diagnosis of *Vibrio cholerae*, Centre for Disease Control and Prevention.

Further tests can be performed to determine the biotype and serotype of the isolate if the specimen tests positive for O1 serogroup. Differentiation can also be undertaken using other phenotypic and genotypic tests. Several well established assays like sheep erythrocyte agglutination, phage, haemolysis, Voges-Proskauer and sensitivity to polymixin B antibiotic can be performed. Genotypic tests can be performed to detect biotype-specific genes such as *tcpA* and *rtxC*. The serotype of O1 isolates can be determined using monoclonal antiserum against the LPS O antgen associated Inaba or Ogawa epitopes. For higher resolution and differentiation between closely related clones of the same biotype, several typing technologies could be used, as discussed in section 1.2.11.

1.2.6 Treatment and Prevention

The main treatment regime of cholera is the rapid rehydration of patients with oral rehydration solutions or intravenous injections of isotonic solutions like WHO approved Ringer's solution in severe cases. Case fatality rate in endemic areas can be reduced to as low as 0.2% just by timely and accurate administration of oral rehydration fluids.

During peak seasons, when hospitals face huge influx of patients, antibiotics are used alongside the rehydration therapy. This is primarily for reducing the diarrheal illness, cutting down the transmission rate and shortening the stay of the patient in the hospital (Lindenbaum, *et al.*, 1967; Sack, *et al.*, 1978). Tetracycline and doxycycline have been long used but other broad range antibiotics are also effective in treatment of severe cholera. With the increased use of antibiotics, cholera strains expressing resistance to the mainline drugs have appeared. For the treatment of these multi-drug resistant strains, ciprofloxacin is recommended. In the case of malnourished children and pregnant women, erythromycin and furazolidone are considered safe (Mandal, *et al.*, 2011). Epidemiological data can play an important role in directing the selection of the right antibiotic. During an outbreak, antibiograms from the public health agency are made available and suitable drug choice should be made in accordance.

Without doubt, the best way of preventing cholera is to improve sanitation and hygiene and make safe drinking water available. European countries that once suffered a burden of cholera in 19th century are an example of the success that could be achieved by adopting these simple measures. However, in parts of the world where cholera is common today, a substantial amount of work, long-term investment and local and government support will be needed to get anywhere close to achieving these goals. In these areas, the next most effective approach is arguably vaccination.

The first vaccine against cholera was developed in the late 19th century and was in use until the 1970s, when it fell out of favour because of limited efficacy, side effects and the injectable mode of administration. This whole cell-based injectable vaccine was largely replaced by a variety of oral killed whole cell vaccines. Dukoral was the first to be marketed and is to date possibly the most successful WHO approved cholera vaccine. It is currently produced and distributed by Crucell and contains the recombinant B-subunit of cholera toxin along with the killed whole *V. cholerae* cells. This vaccine, initially developed by Jan Holmgren and colleagues in Sweden elicits both anti-bacterial serum vibriocidal activity and an anti-toxin immune response (Holmgren, *et al.*, 1977). The vaccine has been shown to provide up to three years of immunity to recipients of two doses in Bangladesh and Peru (Clemens, *et al.*, 1990). It can provide up to 85% protective efficacy and in areas of high vaccine coverage, unvaccinated population can benefit from a herd immunity effect of the vaccination. However, since the price of this vaccine is high due to the recombinant B-subunit and requires two doses for achieving optimum efficacy, it is unsuitable for mass vaccination programs. Currently, it is mainly used by travellers and efforts are under way to develop alternative whole cell-based oral vaccines. For example, another vaccine, Orochol, also produced by Crucell is a live attenuated single dose vaccine that has been shown to provide 79% protective efficacy (Calain, *et al.*, 2004). Derived from a classical *V. cholerae* 0569B, this vaccine strain expresses an immunologically active B-subunit and was shown to be safe and immunogenic in volunteer studies. This vaccine was licenced in some countries but because of limited success across the globe and safety concerns, it is not widely used.

A third vaccine, which is currently only available in Vietnam is going through the WHO accreditation process after reformulation and production in India is Shancol. It was developed locally in Vietnam by Vabiotech in collaboration with the International Vaccine Institute, in Korea but is now produced by WHO approved manufacturer Shantha Biotech in India. In a Vietnamese trial involving 50,000 subjects the vaccine showed 66% efficacy over 10 months (Levine, 1997). Shancol is an oral killed whole cell vaccine and is cheap to produce since it does not have any recombinant B-subunit. A recent trial of this vaccine in Kolkata has proven very successful, raising the hopes that this vaccine will get licenced worldwide and become available for mass vaccination programs in endemic and epidemic settings (Sur, *et al.*, 2009). One main limitation of this vaccine is its liquid formulation, which adds to the transport cost and the other is its two-dose regimen. Efforts are being made to make lyophilized but stable form of this vaccine to cut down the costs even further and make this vaccine available in the most remote of the areas.

After the explosive expansion of cholera in African countries and Haiti in the last few years, vaccination is being promoted as the best way forward alongside the active monitoring efforts. The financial sustainability of a cholera vaccine stockpile, mainly Shancol, is being discussed across various national and international public health committees. Successful planning and execution of vaccination programs would hopefully help in capturing cholera outbreaks in their nascent stages in epidemic suffering areas and more importantly control the disease in traditionally endemic source regions.

1.2.7 Molecular basis of pathogenesis and cholera virulence factors

After the *V. cholerae* bacteria find their way to the gut, motility due to flagella helps them move to the epithelial cells of small intestine. Mucinase enzyme expressed by the bacterium help to penetrate the mucosal layer and TCP, encoded by the vibrio pathogenicity island 1 (VPI-1), then facilitate colonization and attachment of bacterial cells to receptors on the epithelial cells (Butler and Camilli, 2005; Lee, *et al.*, 1999; Silva, *et al.*, 2006; Tacket, *et al.*, 1998; Taylor, *et al.*, 1987). Efficient delivery of cholera toxin (CT) directly onto the epithelial cells takes place to begin the first phases of the molecular pathogenesis pathway of the cholera disease (Figure 1.9).

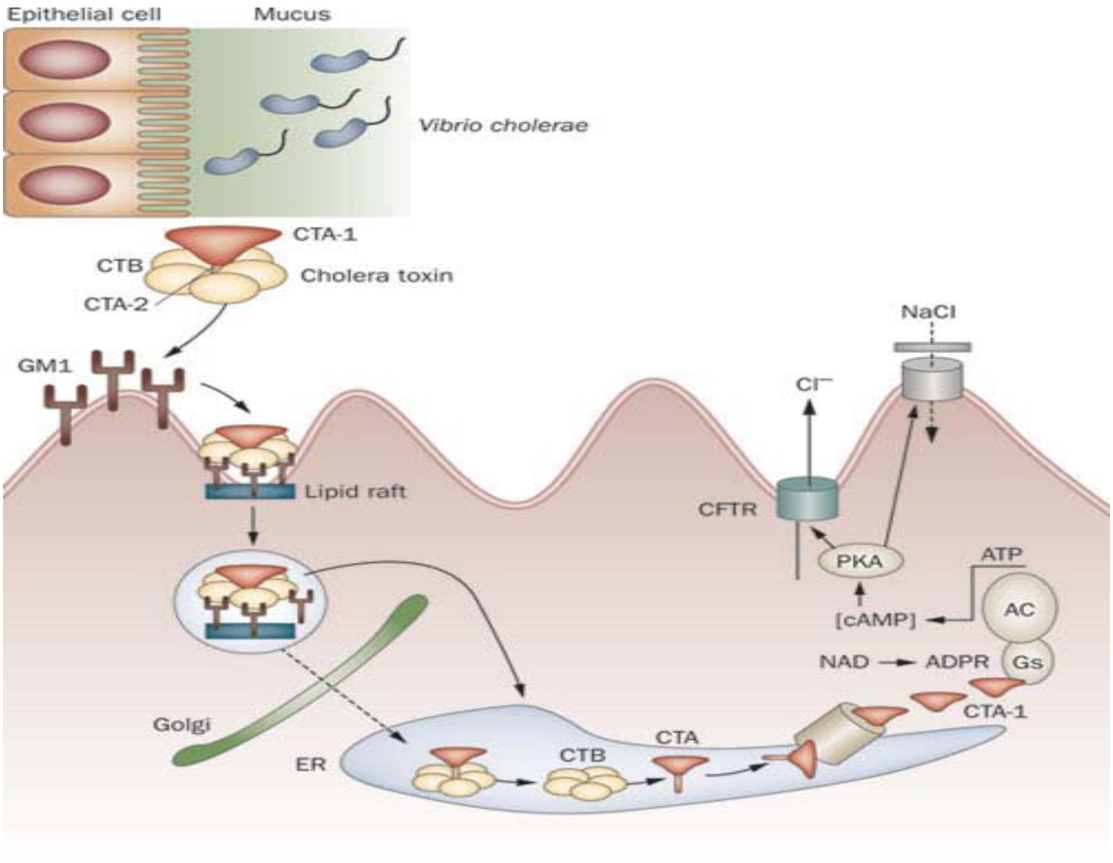

Figure 1.9: Molecular mechanism of the working of CT. Figure reproduced from Clemens *et al.* 2011 (Clemens, *et al.*, 2011).

CT is an AB type toxin consisting of an enzymatic A subunit and pentameric B subunit. The B subunit binds to GM1 ganglioside receptors and the A subunit is endocytosed by the epithelial cells. Once internalized, the A subunit undergoes proteolytic cleavage to release A1 and A2 peptides. The A1 subunit is enzymatically active and catalyzes the ADP ribosylation of the GTP binding G proteins. This activity results in constitutive activation of the adenylate cyclase enzyme, which drives an increase in intracellular cAMP levels. This activity causes excessive secretion of chloride ions into the small intestine and inhibition of sodium chloride absorption, which in turn results in heavy osmotic influx of water from the intravascular spaces of the body into the small intestine and profuse watery diarrhea.

Successful cholera infection, from the bacterial perspective, requires the coordinated functioning of all these and other virulence factors (Butler and Camilli, 2005; Lee, *et al.*, 1999; Silva, *et al.*, 2006; Tacket, *et al.*, 1998; Taylor, *et al.*, 1987) (Figure 1.10).

Astute expression of genes encoding for these virulence factors aids the pathogen in colonize, cause signature disease through toxin production and eventual escape from the intestine, mediating further transmission (Nielsen, *et al.*, 2006; Schild, *et al.*, 2007).

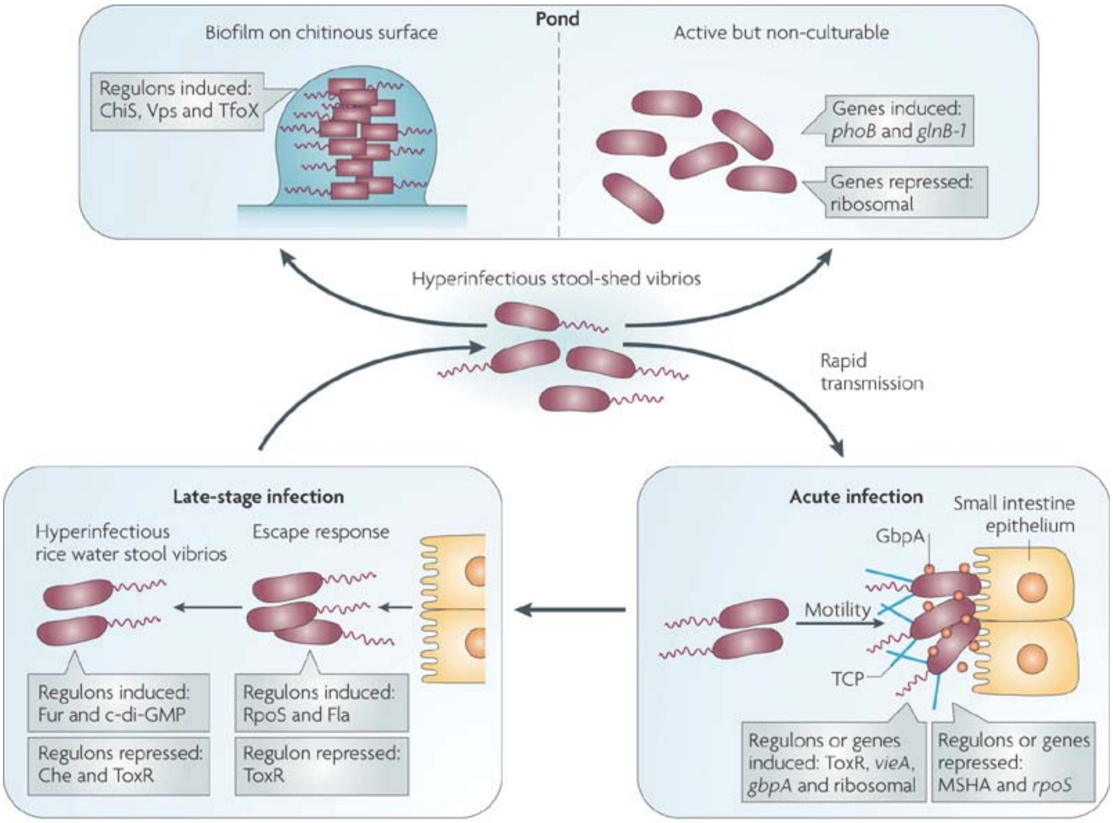
Pond
Biofilm on chitinous surface
Regulons induced: ChiS, Vps and TfoX
Active but non-culturable
Genes induced: *phoB* and *glnB-1*
Genes repressed: ribosomal
Hyperinfectious stool-shed vibrios
Rapid transmission
Late-stage infection
Hyperinfectious rice water stool vibrios
Escape response
Regulons induced: Fur and c-di-GMP
Regulons repressed: Che and ToxR
Regulons induced: RpoS and Fla
Regulon repressed: ToxR
Acute infection
Small intestine epithelium
GbpA
Motility
TCP
Regulons or genes induced: ToxR, *vieA*, *gbpA* and ribosomal
Regulons or genes repressed: MSHA and *rpoS*

Figure 1.10: Expression of virulence factors at different times regulates the establishment of cholera infection and transmission. Figure reproduced from Nelson *et al.* 2009 (Nelson, *et al.*, 2009).

1.2.8 CTX and other *V. cholerae* toxins

Robert Koch, in 1884, proposed that cholera is due to a special poison, which acts on the epithelium, and symptoms of cholera can be termed as poisoning (Koch, 1884). For a substantial period, it was a hypothesis that few believed. It was in 1959 that scientists working in Calcutta demonstrated the existence of a potential poison, which they called enterotoxin (De, 1959). In 1969, efforts to purify the toxin to homogeneity were successful (Finkelstein and LoSpalluto, 1969) and the availability of pure material allowed researchers to discover the toxins detailed biochemical properties and mode of action. In 1983, when 25ul of CT was orally administered to volunteers with sodium bicarbonate solution, many released over 20 litres of rice water stool and the vital role of CT in cholera disease was established (Levine, *et al.*, 1988).

A full structural characterization of CT came after the crystal structure of the highly related labile (LT) toxin of *E. coli* had been determined (Sixma, *et al.*, 1991). The crystal structure CT showed that it is very similar to LT toxin in having a pentameric B subunit and an A subunit (Sixma, *et al.*, 1991). The arrangement is such that the B subunits form a barrel in the center leaving a pore 1.1 – 1.5 nm in size where the A2 peptide of the A subunit sits and binds to the B subunit pentamer. This A2 peptide is linked to the A1 peptide and in whole the A subunit resembles a triangle. This structure is similar to the catalytic region of diphtheria toxin (Sixma, *et al.*, 1991).

The interaction of CT with the GM1 receptors on the intestinal cells occurs *via* the CT B subunit, also sometimes called the choleragenoid because it is a part of choleragen (an alternative name for CT). It was shown that in rabbit ileal loops, if purified B subunit toxin is added before CT, the fluid accumulation is significantly reduced (Pierce, 1973). This gave way to the concept that antibodies against the B subunit are much more protective against cholera than those against the A subunit (Peterson, *et al.*, 1979). Later when the receptors of B subunit binding on the intestinal epithelial cells were recognized as GM1 gangliosides (Holmgren, *et al.*, 1973), it was shown that if GM1 ganglioside is administered into the rabbit ileal loops before CT challenge, the fluid secretion is inhibited (Pierce, 1973).

With the advent of gene cloning and genome sequencing, it was shown that cholera toxin is expressed from two adjacent genes, *ctx*A and *ctx*B, present on an integrated prophage called CTX (for cholera toxin phage) This phage can be integrated as a prophage in the genome of *V. cholerae* or can exist as a replication proficient plasmid (Sack, *et al.*, 2004). Certain conditions can induce toxigenic *V. cholerae* to produce extracellular phage particles (Sack, *et al.*, 2004). Also, non-toxigenic strains can be converted to toxigenic *V. cholerae* on transduction with this phage. It has been proposed that mixed infections that may take place in endemic countries can lead to the generation of new recombinant toxigenic strains by transduction in the gastrointestinal tract (Sack, *et al.*, 2004). The structure and arrangement of genes of a classical CTX phage is shown in Figure 1.11. CTX encodes *ctx*A and *ctx*B, which encode for cholera AB toxin subunits respectively, along with genes encoding the phage structural and regulatory machinery. Other genes carried on CTX include, *psh*, *cep*, *gIII*CTX and *ace*, encoding proteins for phage packaging and secretion and the gene *zot* encodes for phage assembly protein also known as zona occludens toxin. All these genes form the core of CTX phage. Additionally, the genes *rst*R, *rst*A and *rst*B are involved in the regulation of phage secretion and toxin formation and are present on the RS2 region of the classical phage.

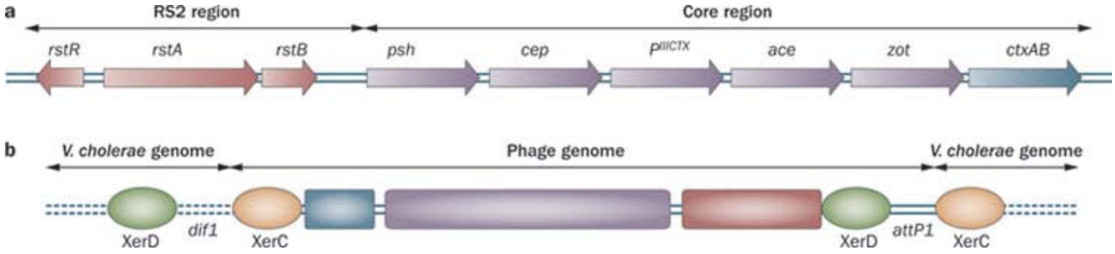

Figure 1.11: a) Arrangement of core and RS2 genes of cholera toxin phage and b) the integration sites in the *V. cholerae* chromosome. Figure reproduced from Clemens *et al.* 2011 (Clemens, *et al.*, 2011).

In conventional seventh pandemic El Tor biotype strains, an additional satellite phage is present, which is the same as RS2 but has an additional gene, *rst*C. The product of *rst*C is a repressor of *rst*R that stimulates El Tor strains to produce more CT. More variants with novel gene arrangements of El Tor CTX have now been discovered and the genomes of some of these are illustrated in Figure 1.12. The most commonly reported of these are the so called atypical and hybrid variants but since CTX is a phage and is mobile, many different arrangements of these genes are possible and it will not be surprising to find several new genome arrangements appearing in the future (as explained in section 1.2.11).

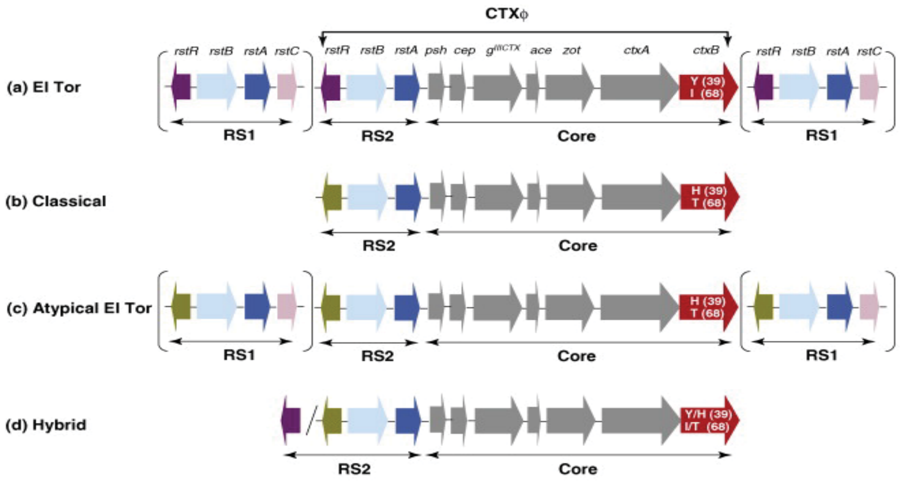

Figure 1.12: CTX phage from classical (a) and El Tor (b) strains have been reported in original respective biotypes but reports of variants of El Tor (c and d) have become more common in recent years. Figure reproduced from Safa *et al.* 2010 (Safa, *et al.*, 2010).

There are also genes in the *V. cholerae* chromosomal back bone that encode helper enzymes that, when expressed, can increase the impact of CT. *nan*H encodes a neuraminidase enzyme NANase that can catalyse the conversion of normal gangliosides to GM1 thereby increasing the chances of CT binding with the intestinal epithelial cells and inducing more fluid secretion (Holmgren, *et al.*, 1975). *dsb*A encodes a disulfide isomerase that catalyses the formation of crucial disulfide bonds between the peptides of A subunit and B subunits (Peek and Taylor, 1992). Thirdly, a gene called *hap* encodes a haemaglutinin/protease, which likely plays a vital role in dissociation of A1 peptide from A2 peptide during cholera pathogenesis. Finally, *V. cholerae* can coordinately regulate the activation and inactivation of the genes that encode for colonization factors and toxins. The *tox*R gene of the ToxR regulon encodes for a protein that can bind to a 7bp sequence found upstream of *ctx*AB and can regulate cholera toxin production in concordance with the environmental or host conditions. The *tox*R gene also regulates the expression of *tox*T gene, which in turn regulates up to 17 genes that form the ToxR regulon (Parsot and Mekalanos, 1990; Skorupski and Taylor, 1997).

V. cholerae can produce toxins other than CT and CTX-negative *V. cholerae* can cause mild diarrhea (Levine, *et al.*, 1988). Volunteers ingesting 10^4 to 10^{10} non-toxigenic *V. cholerae* experienced moderate diarrhea that lasted for 3 days and up to 2 liters of stool was released. A gene called *hly*A encodes for hemolysin and is present in all *V. cholerae* and this protein has been shown to cause fluid secretion when injected in rabbit illeal loops (Ichinose, *et al.*, 1987). The fluid however is different from that produced in response to CT, since it is bloody and contains mucus (Ichinose, *et al.*, 1987). Some researchers have proposed that *zot* and *ace* also encode for products with enterotoxic activity as they increase short circuit current in rabbit intestines in Ussing chambers (Fasano, *et al.*, 1991; Trucksis, *et al.*, 2000).

1.2.9 Vibrio pathogenicity and seventh pandemic islands

V. cholerae strains of both classical and El Tor biotype possess vibrio pathogenicity islands 1 and 2 (VPI 1 and 2). While these two islands are useful genetic markers for epidemic causing strains, vibrio seventh pandemic islands 1 and 2 (VSP-1 and 2) are only present in the seventh pandemic lineage of *V. cholerae*. Although cholera toxin can cause severe disease when orally administered by itself, some genes of VPI-1 islands are essential for *V. cholerae* to establish an infection.

VPI-1 encodes for the toxin co-regulated pilus (TCP), which is fundamental in bacterial colonization of the intestine (Sack, *et al.*, 2004). *V. cholerae* strains lacking this island fail to attach to the surface of the intestinal epithelial cells and are quickly cleared by the foreign antigen cleansing mechanism of the small bowel. Several studies have suggested that this island (40kb in length) is derived from a bacteriophage (Sack, *et al.*, 2004) and genes in the TCP operon encode for virulence, regulation and a transposase. At either end of the island there are attachment sites for site-specific integration into the genome of *V. cholerae* (Sack, *et al.*, 2004). The arrangement of genes of the TCP cluster is shown in Figure 1.13.

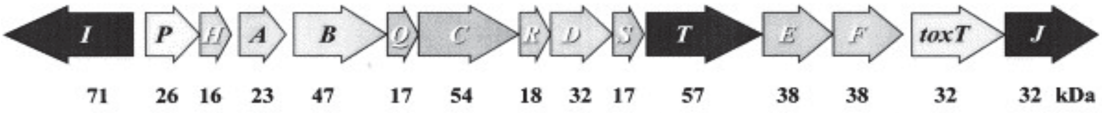

Figure 1.13: The TCP gene cluster of *V. cholerae*. Figure reproduced from Manning

1997 (Manning, 1997).

VPI-2 is a 57 kb island on the *V. cholerae* genome. Its GC content (42%), the presence of an integrase, and being inserted next to tRNA in a region flanked by direct repeats is indicative of a bacteriophage origin and being acquired by horizontal gene transfer (Jermyn and Boyd, 2002). The genes present on this island encode for sialic acid transport and catabolism machinery alongside a neuraminidase, which is a helper enzyme for more effective action of CT. The neuraminidase enzyme also forms part of the mucinase complex that breaks up the intestinal mucosal layer and helps bacteria to penetrate to the site of attachment and CT action. The arrangement of VPI-2 genes is shown in Figure 1.14.

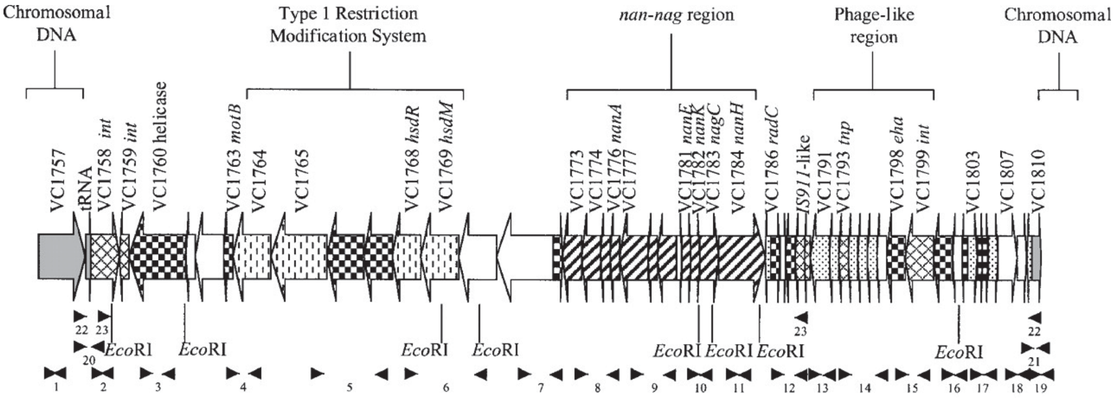

Figure 1.14: Arrangement of all the genes on the VPI-2 island, important regions are marked by their CDS number in the *V. cholerae* genome. Figure reproduced from Jermyn *et al.* 2002 (Jermyn and Boyd, 2002).

VSP-1 is a 16-kb island inserted in the *V. cholerae* genome that encodes 11 CDSs (VC0175-VC0185) in the *V. cholerae* El Tor genome. It has a GC content of 40%, which is different from the 47% of the whole genome backbone, suggesting that it has been horizontally acquired. Full phenotypic characterization of the genes in this island is yet to be reported but the di-nucleotide cyclase enzyme encoded on this island promotes colonization and plays a role in *V. cholerae* chemotaxis (Davies, *et al.*, 2012). VSP-2 is 27kb long, is integrated at a tRNA and possesses a P4-phage like integrase. It constitutes CDSs VC0490 to VC0516 on the *V. cholerae* O1 genome with genes encoding for RNase, a type IV pilus, a DNA repair protein, two transcriptional regulators and two methyl-accepting chemotaxis proteins (Davies, *et al.*, 2012). The exact function of this island is also unknown and many strains of the seventh pandemic lineage now have variants of VSP-2, where other genes have replaced parts of it by homologous recombination.

1.2.10 Multiple antibiotic resistance cassettes

In addition to virulence or pathogenicity, some *V. cholerae* encode genomic islands associated with multiple antibiotic resistance determinants. Antibiotic resistant *V. cholerae* strains were first reported in Tanzania in 1977 and later in Bangladesh. Multiple antibiotic resistance cassettes harbored by integrons are the main drivers of resistance in *V. cholerae*. These integrons are either a part of integrative conjugative element (ICE) or super-integron.

SXT (denoting sulfamethoxazole-trimethoprim) is the ICE element of *V. cholerae* that was first identified in 1992 in the then newly discovered serogroup O139 strains (1993). This island is ~100kb in size and encodes resistance to multiple antibiotics including sulfamethoxazole and trimethoprim (which give the element its name). After its discovery, SXT or its variant ICE form has been found in many seventh pandemic O1 El Tor strains and even in genera outside *Vibrio* (Ahmed, *et al.*, 2005), proving the horizontally transferrable nature of these elements (as discussed below). They integrate into the host chromosome at a specific site, in the *prf*C gene, and can excise perfectly without leaving any scar. Their excision is such that the gene they disrupt during integration is reformed and its activity is totally resumed. SXT are genetically closely related to the IncJ element of the R391 family of plasmids (Hochhut, *et al.*, 2001), but now it has become clear that IncJ plasmids are actually ICE elements that can integrate into the genome and excise to facilitate horizontal transfer to a variety of other Gram-negative bacteria (Waldor, *et al.*, 1996).

The structure of SXT, its core genes and hot spots have been best described in the detailed work of Wozniak *et al.* (Wozniak, *et al.*, 2009) (Figure 1.15). These researchers described the ICE elements of *Photobacterium damselae, Shewanella putrefaciens, Providencia rettgeri* and several strains of *V. cholerae* and found that all the ICE elements belonged to the R391 ICE family and differed only in the integron cassettes inserted at different hot spots.

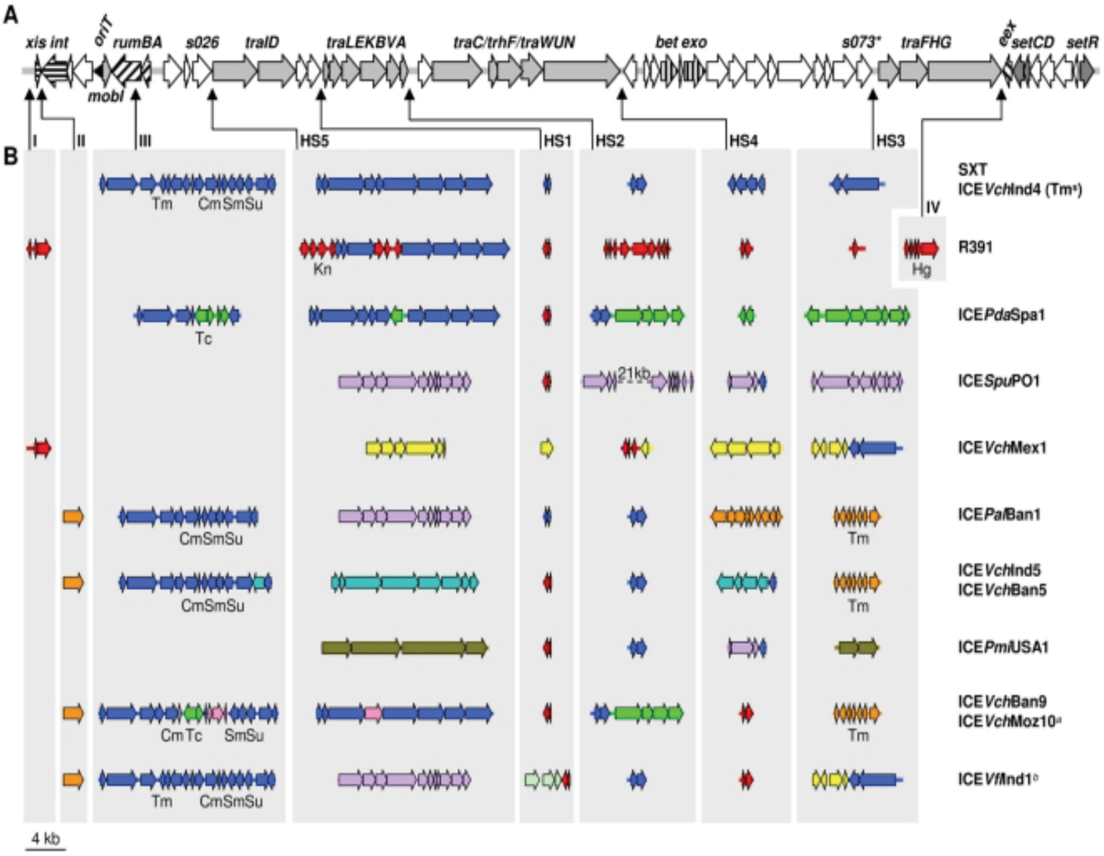

Figure 1.15: The structure of SXT/R391 family of ICE elements and the hot spot regions where different antibiotic resistance gene cassettes are inserted. Figure reproduced from Wozniak *et al.* 2009 (Wozniak, *et al.*, 2009).

Resistance against some antibiotics is driven by mutations in chromosomal genes. For instance, specific mutations in *gyr*A and *par*C genes can convert strains normally susceptible to quinolones into a resistant form (Mukhopadhyay, *et al.*, 1998). In addition, proton motive force driven efflux pumps can make strains more resistant to these antibiotics (Baranwal, *et al.*, 2002).

1.2.11 Typing schemes for *V. cholerae*

While phenotypic identification and confirmation has its merits in hospital environments, rapid genotypic identification of strain types is important from a public health perspective. For understanding any outbreak, links between strains need to be established and therefore many typing techniques are now available for identifying the various strain types of *V. cholerae*. One of the first and most common typing techniques is CTX typing. The different gene arrangements in classical and El Tor CTX phage and in addition the amino acid differences in the proteins encoded by the genes *rst*R and *ctx*B are used for typing strains into either the classical and El Tor biotype. *rst*C gene, which is a part of the RS1 satellite phage, is normally only present in the El Tor strains. When atypical and hybrid variants were found, this typing scheme had to be expanded to include variants such as El Tor CTX structures with *ctx*B of classical became predominant. However, when multiple genetic arrangements of CTX phage genes started appearing, as illustrated in Figure 1.16 (Chun, *et al.*, 2009), researchers started considering this typing as near obsolete.

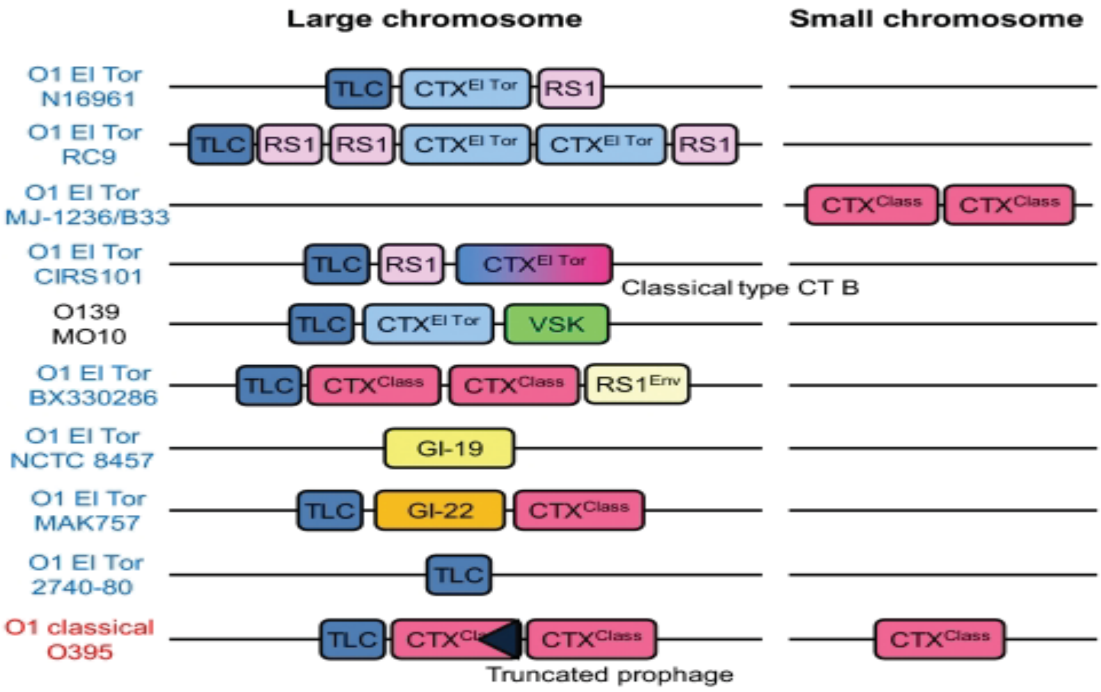

Figure 1.16: Many possible arrangements of genes within different CTX phages from different *V. cholerae* strains. Figure reproduced from Chun *et al.* 2009 (Chun, *et al.*, 2009).

CTX is a mobile phage and therefore typing based on this mobile element has limited utility. Multi locus genotyping starting with MLEE (multi locus enzyme electrophoresis) was in use from an early stage. With the improvement in sequencing technology and cloning, MLEE evolved into MLST (multi locus sequence tag) and currently the technique most in use is MLVA (multi locus variable number tandem repeat analysis). However, all these techniques look at a very limited number of loci and lack the phylogenetic context.

PFGE (pulse field gel electrophoresis) is currently the gold standard of *V. cholerae* genotyping in public health laboratories worldwide. It uses a mixture of restriction enzymes to cut the *V. cholerae* genome at multiple locations followed by an overnight run of the restricted DNA on agarose gel to obtain a DNA fingerprint. The pattern obtained is matched between the strains and is used to group various clones together. Although robust, this technique is laborious and it does not highlight the much needed phylogenetic relationship between the strains that fall in different PFGE pattern groups.

The advent of whole genome sequencing has facilitated studying of bacterial genomes and diversity between the strains in detail at a level that was not possible before this technology was developed. Differences at single base pair levels can now be monitored to build robust phylogenetic trees and understand family history of strains.

1.3 Whole genome sequencing

It can be argued that the path to whole genome sequencing began with the development by Frederick Sanger and his collaborators of the chain termination sequencing method in 1970s (Sanger and Coulson, 1975). This technology was the gold standard until less than ten years ago and is still in use. The first complete DNA genome to be sequenced was that of bacteriophage phix174 in 1977 (Sanger, *et al.*, 1977). The automated version of Sanger technique brought the direct use of computers in sequence analysis. The first bacterial genome to be sequenced was that of *H. influenzae* in 1995 (Fleischmann, *et al.*, 1995) but it was the landmark publication of human genome in 2004 (2004) that proved the worth of sequencing and opened up gateways to high throughput sequencing of a variety of organisms. To deliver at this scale, even the automated capillary sequencers were not enough and it was the advent of the next generation sequencing technologies that transformed the world of sequencing.

1.3.1 Next-Generation sequencing

The numbers of completed genomes and projects currently in progress have exponentially increased in the genome online database (GOLD; http://www.genomesonline.org/). Next generation sequencing, a term used to refer to all new technologies developed after the Sanger sequencing, has several advantages over the later. First, it does not involve any cloning step and therefore reduces the cost and time of sequencing. Second, it allows sequencing of many DNA fragments in parallel (Shendure and Ji, 2008). Third, each base is sequenced multiple times (referred to as coverage), which reduces the number of false positive calls. The cons on the other hand lie in shorter read lengths and assembly challenges. However, innovative approaches have been invented to tackle these issues and make the best use of the big datasets that the Next-Gen sequencers provide (Pop and Salzberg, 2008). A bacterial genome that used to take years to finish by Sanger sequencing can now be sequenced very rapidly.

1.3.1.1 New sequencing technologies

The "Next-Generation (Next-Gen)" sequencing technologies that are currently at the front end are 454 (Roche), Genome Analyser II or Hi Seq (Illumina/Solexa) and SOLiD (Applied Biosystems). The working platform for all these technologies is similar. All involve random fragmentation of genomic DNA, amplification directly on the surface (bead/chip) and use of powerful camera optics to record the base incorporated.

454 technology uses sequencing by synthesis methodology (Margulies, *et al.*, 2005). DNA fragments are attached to the beads and are amplified by emulsion PCR. DNA carrying beads are then loaded onto the pico titre plate with tiny wells and sequencing reagents flow onto the plate. The addition of a new base results in the release of a pyrophosphate and a chemical reaction converts luciferin to oxy-luciferin and light. This light is picked up by the camera and base calls are made (Margulies, *et al.*, 2005). Though this technology gives longer reads compared to other Next-Gen technologies and therefore is a preferred choice for de-novo assemblies, it can mis-predict the length of homo-polymeric sequences. Since only one type of nucleotide

can be added at a time, addition of multiple identical bases normally measured by light intensity become difficult to judge when homo-polymers are longer than a certain length (Shendure and Ji, 2008).

The SOLiD platform of Applied Biosystems, on the other hand, is based on sequencing by ligation methodology (Shendure and Ji, 2008). Amplification takes place in water-oil emulsion as in 454 technology, however the sequencing chemistry is very different. A probe is ligated to the DNA carrying bead and once a base is incorporated, the image is captured by the camera. The probe is finally cut and washed off before the same cycle is repeated to note the sequence.

From costs per run and costs per gigabyte of data perspective, Illumina sequencing is currently the leader (Liu, *et al.*, 2012). This technology works on a sequencing by synthesis basis but the difference is in the surface on which DNA fragments are attached for amplification and sequencing. It uses a flat chip called a “flow cell” instead of bead and amplification takes place on the surface by bridge PCR. Multiple identical or complementary fragments generated by this amplification cycle are sequenced. This is achieved by flowing all four types of reversible di-deoxy-nucleotides onto the flow cell surface and monitoring each base incorporated by means of image capture. Several studies have suggested cons in this technology too (Dohm, *et al.*, 2008; Harismendy, *et al.*, 2009; Quail, *et al.*, 2008) but improvements have also been proposed to minimize the negatives (Quail, *et al.*, 2008).

1.3.1.2 Next-Generation bioinformatics tools

Next Generation technologies have made sequencing truly high throughput, but the analysis of short read data generated its own new challenges that had to be dealt with (Pop and Salzberg, 2008). For example, the *de novo* assembly of 100-400 bp reads and mapping of data (with multiple coverage for each base) to call the variants required new strategies. Genome assembly of these short reads is performed by identifying overlaps in short reads and joining them to form a contig. When paired end sequencing is achieved, read pair information can be used to further join the contigs into “super contigs” or “scaffolds”. However, in the regions with repeats and low coverage, short reads fail to assemble properly (Miller, *et al.*, 2010).

For *de novo* assemblies, capillary data still has no match but 454 data provides the best read-length amongst all the Next-Gen outputs. Although Pacific BioSciences' third generation single molecule real time (SMRT) sequencing (Korlach, *et al.*, 2010) provides an average of 5000 bp reads, the assemblies without the data from other technologies lack robustness because of its high base calling error rate (Koren, *et al.*, 2012). Studies have shown that some parametric optimizations can give good assemblies of Illumina data too (Hernandez, *et al.*, 2008; Studholme, *et al.*, 2009). There are several assembly software that have been written to work with large datasets like those from Next-Gen sequencers (Li, *et al.*, 2010; Zerbino and Birney, 2008). The most used is Velvet assembler of Zerbino and Birney (Zerbino and Birney, 2008), which also takes the read pair information into account when the data is in paired end read format. All short read data assemblers give N50 values, which indicate the quality of *de-novo* assemblies. N50 value is the length of the smallest contig in the scaffold set that contains the fewest and therefore the largest contigs, which in total length represent at least 50% of the assembly (Miller, *et al.*, 2010).

There are assemblers like "Celera", which can use data of multiple formats and form the best assembly (Pop and Salzberg, 2008). It utilizes the longer read data to fill the gaps left between contigs or scaffolds. Use of mixed platform data for *de novo* assemblies has been shown to give best assemblies with high N50 values (Aury, *et al.*, 2008). Since assemblers take coverage, read pairs and base quality scores into account, changing the parameters of assembler runs can affect the N50 values. Now, there are assemblers like Velvet Optimiser (www.bioinformatics.net.au/software.velvetoptimiser.shtml), which are freely available and can optimize the parameters to fit the data quality to output best assemblies. Due to the volume of data being generated by these Next-Gen sequencing technologies, it is impossible to completely finish every assembly into a finished genome. Scientists today use near accurate assemblies or "draft genomes" to look for genomic regions of differences. Although, calls cannot be made in repetitive regions or regions with low coverage because of lack of statistical confidence, the parts of genome that are completely assembled into a contig can be easily looked for insertions, deletions or recombinations (Chain, *et al.*, 2009).

From the point of view of public health reference labs, epidemiologists and outbreak monitoring agencies, variation detection is more important than the whole genome sequence assembly of any new organism. Next generation sequencing provides the highest resolution data currently possible and single base pair level polymorphisms (SNPs) can be detected by mapping raw data to a reference (completed) genome. There are several mapping programs like MAQ, BWA, SSAHA, Bowtie and SMALT, freely available for the research community (Langmead, *et al.*, 2009; Li and Durbin, 2010; Li, *et al.*, 2008; Ning, *et al.*, 2001). They take into account the read length, shape of the reference genome, and base quality score to produce best possible alignments to the reference.

1.3.2 Understanding bacterial evolution and transmission using genomics

Next generation sequencing has transformed the world of bacterial genotyping. Since bacterial genomes are predominantly much smaller than eukaryotic or mammalian genomes, multiple genomes can be sequenced in one run of Illumina/454/SOLiD machines. Since Illumina provides the least error prone and highest volume of data, literature search shows that this is the most popular platform of choice. The generated sequence data can be used in a variety of studies from bacterial evolutionary genetics, comparative genomics, transcriptomics, outbreak tracking, metagenomics and transmission studies.

The resolution provided by Next-Gen data has allowed studies on the population structure of even the most clonal bacterial populations. Traditional typing technologies were not able to distinguish between these monomorphic pathogens and therefore confirming transmission between individual patients or sometimes within a geographical boundary was not possible. Several studies are now available that illustrate the huge potential of genomics based variation detection in public health. Noticeable examples include studies on *Staphylococcus aureus* (Harris, *et al.*, 2008), *Streptococcus pneumoniae* (Croucher, *et al.*, 2011), *Salmonella Typhimurium* (Okoro, *et al.*, 2012), *Mycobacterium tuberculosis (*Bryant, *et al., 2013), Chlamydia trachomatis* (Harris, *et al.*, 2012), *Shigella sonnei* (Holt, *et al.*, 2012) *and Clostridium difficile* (He, *et al.*, 2010) among others (Chin, *et al.*, 2011; Chun, *et al.*, 2009; Hendriksen, *et al.*, 2011). Metagenomics and studies looking at the spread of antibiotic resistance in bacteria and parasites have also utilized the power of extensive data and deep sequencing to identify newly emerged pathogenic clades (Adler, *et al.*, 2013; Holden, *et al.*, 2013; Miotto, *et al.*, 2013).

Researchers investigating basic biology and microbiology have also made novel use of these sequencing technologies. Techniques like transposon dependent insertion sequencing (TraDIS or Tn-Seq) have facilitated the simultaneous sequencing of libraries harbouring more than a million transposon mutants within a single strain of a pathogen. These mutants can be screened to identify genes that are essential for survival under certain conditions or contribute to a particular phenotype (Barquist, *et al.*, 2013; Langridge, *et al.*, 2009; van Opijnen and Camilli, 2013). Sequencing of cDNA/RNA can be performed (RNA-Seq) to explore the transcriptome of pathogen or host, for example to investigate differential expression of genes in the infection life cycle *in vivo* and *in vitro* (Albrecht, *et al.*, 2010; Perkins, *et al.*, 2009; Sharma, *et al.*, 2010; Tanaka, *et al.*, 2013).

1.4 *V. cholerae* genomics and genetic diversity

The first *V. cholerae* genome to be completely sequenced was that of the seventh pandemic O1 El Tor strain N16961, isolated in 1975 in Bangladesh. This genome was sequenced on capillary machines by a whole genome random sequencing method and manual assembly (Heidelberg, *et al.*, 2000). The *V. cholerae* genome is unusual and is different from many other Gram-negative bacteria because it incorporates two independently replicating circular chromosomes (Figure 1.17).

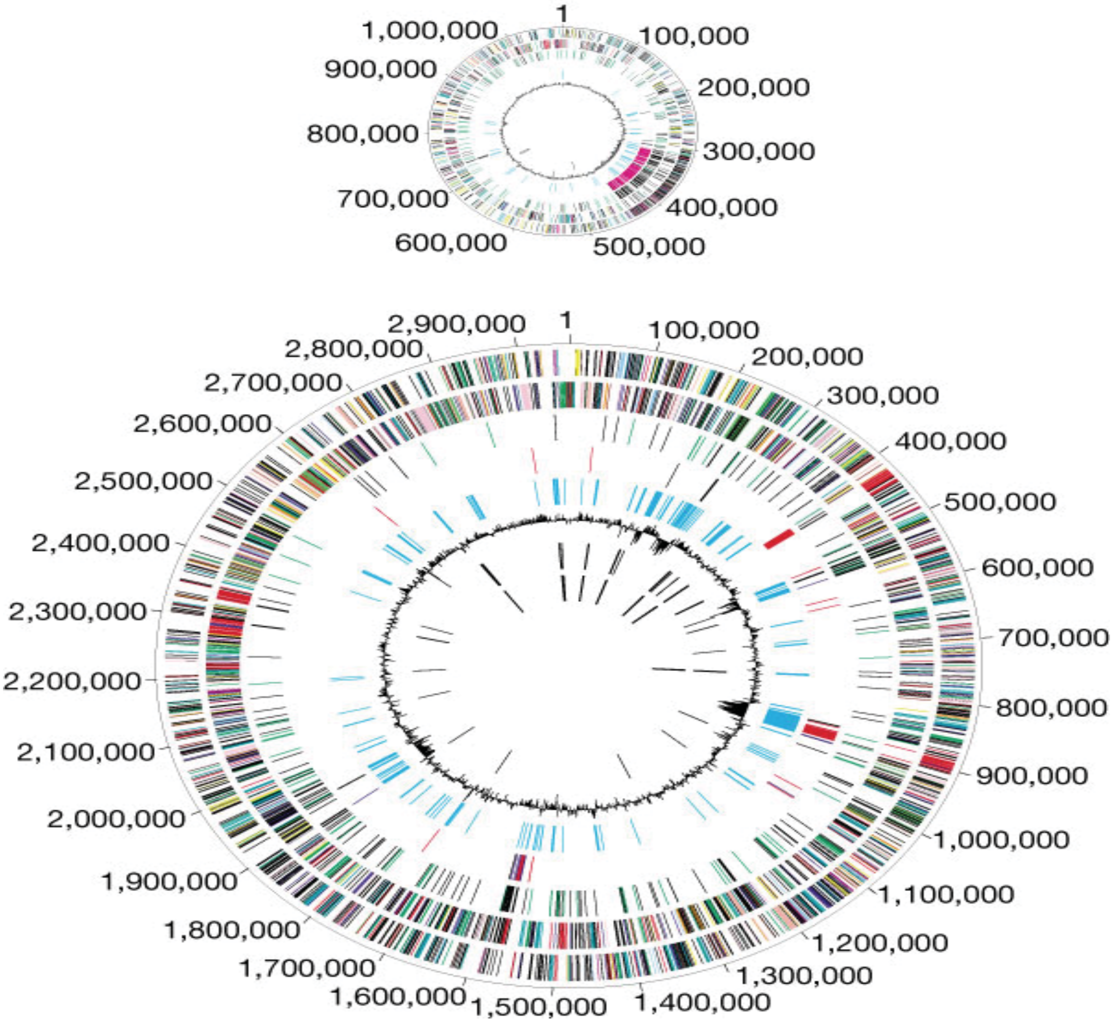

Figure 1.17: The genome of *V. cholerae* incorporates a ~1 Mb chromosome 2 (top) and ~3 Mb chromosome 1 (bottom) (Heidelberg, *et al.*, 2000). See original reference for full description of this figure.

While most of the housekeeping (e.g. DNA replication, transcription, cell wall synthesis and translation) and pathogenicity (e.g. toxins, colonization factors, toxin regulatory genes and LPS surface antigens) genes are located on chromosome 1, chromosome 2 contains a large number of hypothetical genes and a super-integron of 173 integron cassettes that covers a considerable length of this chromosome (Heidelberg, *et al.*, 2000). It is believed that a significant proportion of the small chromosome may have originally entered Vibrio as a mega plasmid because some genes on this chromosome are normally found on plasmids. The average GC contents of chromosome 1 and chromosome 2 are 46.9 % and 47.7 % respectively.

Approximately 1450 genes on both chromosomes of *V. cholerae* are similar to genes present in *E. coli* genomes, but approximately 500 of these represent potential gene duplications (Heidelberg, *et al.*, 2000). These genes mostly encode for products involved in regulatory functions, chemotaxis, pathogenicity and transport. Since *V. cholerae* is naturally an aquatic bacterium, the presence of multiple copies of genes involved in chemotaxis, nutrient transport and quorum sensing are perhaps not surprising.

The complete genome of N16961 is the most widely used reference genome for comparative genomics and evolutionary studies of seventh pandemic *V. cholerae*. However, several insertion and deletions of the genome that were not in the original publication have been identified by Andrew Camilli's group in USA (personal communication) and we incorporated these in the reference sequence before carrying out our analyses.

1.5 Aims and objectives of this study

The main aim of the work described in this thesis was to understand the global and regional level evolution of *V. cholerae* utilizing the fine resolution provided by whole genome data. To begin to define the differences in the genomes of environmental and epidemic *V. cholerae*, we gathered a global collection from all the inhabited continents where cholera is ripe today. Overall, we have analysed the data from over 1000 seventh pandemic and ~50 environmental isolates in our collection of sequences and have mined previously and retrospectively published genomes to construct phylogenies and to begin to understand the evolutionary relationships between them. In cases where detailed sample information or meta-data was available, we used phylogeny alongside clinical, phenotypic, geographical or other meta-data to track and understand the global and regional spread of cholera.

Chapter 2 describes the genomic variation and phylogenetic patterns we identified in *V. cholerae* samples collected for over a century. This analysis highlights the geographical clustering of isolates and the clock-like evolution of the seventh pandemic *V. cholerae*. The role of lineage specific markers and recombination was elucidated and a framework for accurate determination of future epidemics within the seventh pandemic was constructed.

Chapter 3 focuses on subsets of isolates within the global collection and investigates microevolution within geographical areas or national boundaries. First, we show that the molecular clock rate was consistent in a collection of related *V. cholerae* from Pakistan that were collected during the 2010 floods. Second, a surveillance study provides details of the phylogenetic lineages causing cholera in Kenya and shows the limitations of MLVA, one of the most used typing techniques for *V. cholerae*. Finally, a study on Mexican *V. cholerae* populations identifies novel non-CT toxin encoding local epidemic lineages that are unique to that region of the world.

The work described in Chapter 4 investigates a collection of *V. cholerae* isolated during a phase-three vaccine trial in Kolkata, India. The dataset provides clear evidence of serotype switching from one year to the other and details various mutations that could lead to conversion of wild type Ogawa serotype to the mutant Inaba serotype.

Chapter 5 describes the potential use of genomic data, the phylogenetic framework and single base variations for designing a SNP genotyping scheme. We designed two kits for detecting these SNPs and researchers would be able to use their kit of choice depending upon the resolution required.

Chapter 6 concludes the thesis by discussing the future implications of this work and the public health lessons that could be learnt from similar studies.