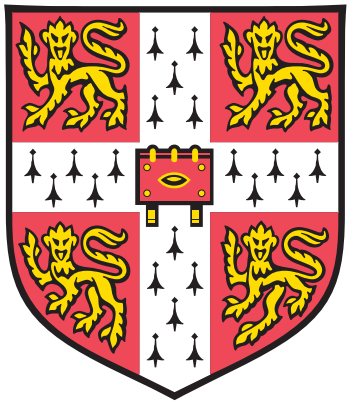
The Origins and Evolution of *Vibrio cholerae* O1 El Tor



Ankur Mutreja

Corpus Christi College

University of Cambridge

This dissertation is submitted for the degree of
Doctor of Philosophy

August 2013

Declaration

This dissertation describes my work undertaken at the Wellcome Trust Sanger Institute between May 2010 and August 2013, under the supervision of Profs. Gordon Dougan, Nicholas Thomson and Julian Parkhill in fulfillment of the requirements for the degree of Doctor of Philosophy, at Corpus Christi College, University of Cambridge.

This thesis is the result of my own work and where the work done in collaboration is presented, it is indicated in the text. The work described here has not been submitted for a degree, diploma or any other qualification at any other university or institution. I confirm that this dissertation does not exceed the page limit specified by the Biology Degree Committee.

Ankur Mutreja

Cambridge, August 2013

Publications

Mutreja, A., Kim, D.W., Thomson, N., *et al*. (2011) Evidence for multiple waves of global transmission within the seventh cholera pandemic, *Nature*, **477(7365)**, 462-465.

Mutreja, A. (2012) Bacterial frequent flyers, *Nature Reviews Microbiology,* **10(11)**, 734.

Kiiru, J.N., Mutreja, A., *et al.* (2013) A study on the geophylogeny of clinical and environmental *Vibrio cholerae* in Kenya, *PLoS ONE,* **8(9),** e74829.

Ali, M., Mutreja, A., *et al.* (2014) Genomic Epidemiology of Vibrio cholerae O1 associated with Floods, Pakistan, 2010, *Emerging Infectious Diseases,* **20(1)**, 13-20.

Acknowledgements

I would like to sincerely thank my supervisors, Profs. Gordon Dougan, Nicholas Thomson and Julian Parkhill for allowing me to do this project, and for guiding, advising, and encouraging me throughout my PhD. I am very grateful to my thesis committee advisors Drs. Julian Rayner and Matt Berrimen, and Prof. James Wood for incredibly useful discussions and constructive criticism. My utmost gratitude goes to the Wellcome Trust for funding my research and maintenance costs.

I am very thankful to all our collaborators around the world who contributed to the *V. cholerae* collections that are at the core of this PhD. Jan Holmgren, Michael Lebens, Sam Kariuki, John Clemens, Alejandro Cravioto, Dong Wook Kim, Habib Bukhari, Cecil Czerkinsky and GB Nair played huge roles in making sure that representative samples were collected from the cholera affected sites and shipped to the WTSI .

I am extremely grateful to WTSI sequencing and pipeline teams: Michael Quail, Richard Rance, David Harris, Elizabeth Gibson, Craig and Nicola Corton, Hilary Browne, Graham Rose, Karen Brooks, Christine Burrows, Louise Clark, Vicky Murray, Scott Thurston, Andries van Tonder, and Danielle Walker have all done a tremendous job in generating the sequencing data used for this analysis. I would also like to thank Jacqui Keane and her team for providing the troubleshooting help in solving informatics related problems.

I am very grateful to Maria Fookes, who kindly helped me with the phylogenetic analyses during my rotation on this project and provided creative ideas when I first started. I am thankful to Simon Harris for sharing his expertise with me and helping me analyse results, especially during the early days of my project. Also, I would like to convey my thanks to Tom Connor for sharing his knowledge of Bayesian analysis software.

I also want to thank Sophie Palmer, Theresa Feltwell, Annabel Smith and Christina Hedberg- Delouka for looking after me at all the steps of my PhD and keeping checks that I was progressing well. I would also like to thank the travel office team, Jeanne

Cook and Anne Wombwell for their cooperation in arranging travel to important meetings and very useful conferences during my PhD. I wouldn't be here without the support of my colleagues and friends, Ravi Verma, Gaurav Godara, Deepak Singh Rana, Deepak Agrawal, Abhinav Prasad and Popoola Olalekan among several others whose words and presence kept me positive at all times.

Dedicated to

My Parents and Sister

"Always appreciate what you have and what you are getting from life, otherwise you risk losing it. Be very proud of yourself."

Contents

The Origins and Evolution of *Vibrio cholerae* O1 El Tor i
Declaration ii
Publications iii
Acknowledgements iv
Dedicated to vi
Contents vii
Figures x
Tables xii
Abstract xiii

1. Introduction 1
1.1. Cholera 1
1.1.1. Overview 1
1.1.2. Cholera Pandemics 2
1.2. *Vibrio* bacteria 5
1.2.1. *V. cholerae*: the species and classification 6
1.2.2. Ecology of *V. cholerae* 9
1.2.3. Epidemiology 11
1.2.4. *V. cholerae* infection and symptoms 13
1.2.5. Diagnosis 14
1.2.6. Treatment and prevention 16
1.2.7. Molecular basis of pathogenesis and cholera virulence factors 18
1.2.8. CTX and other *V. cholerae* toxins 20
1.2.9. *Vibrio* pathogenicity and seventh pandemic islands 24
1.2.10. Multiple antibiotic resistance cassettes 26
1.2.11. Typing schemes for *V. cholerae* 27
1.3. Whole genome sequencing 29
1.3.1. Next generation sequencing 30
1.3.1.1. New sequencing technologies 30
1.3.1.2. Next-Generation bioinformatics tools 31
1.3.2. Understanding bacterial evolution and transmission using genomics 33
1.4. *V. cholerae* genomics and genetic diversity 34
1.5. Aims and objectives of this study 36

2. Genomic variation in global *V. cholerae* spanning a century 38
2.1. Introduction 38
2.2. Bacterial isolates 40
2.3. Results and discussion 44
2.3.1. Global phylogeny of the *V. cholerae* species 44
2.3.2. Evolution of the seventh pandemic O1 El Tor *V. cholerae* 47
2.3.3. The three waves of seventh pandemic O1 El Tor *V. cholerae* 50
2.3.4. The origins of O139 serogroup strains 53
2.3.5. Evidence within the global phylogeny of intercontinental transmission 54
2.3.6. Patterns of gene acquisition and loss in the seventh pandemic 54
2.3.7. Variations in CTX and their phylogenetic distribution 55

2.3.8. Variations in SXT and its phylogenetic distribution 58
2.3.9. WASA-1 and other markers of the West Africa/South American (WASA) clade 63
2.3.10. Recombination 67
2.4. Conclusion and lessons from global phylogeny 68

3. Case studies on the regional evolution of *V. cholerae* O1 El Tor 69
3.1. Introduction 69
3.2. Bacterial isolates 72
3.2.1. Pakistan *V. cholerae* collection 72
3.2.2. Kenyan *V. cholerae* collection 73
3.2.3. Mexican *V. cholerae* collection 76
3.3. Results and discussion 78
3.3.1. Whole genome phylogeny of 2010 Pakistan flood *V. cholerae* 78
3.3.2. Evidence for a strict *V. cholerae* molecular clock in Pakistan 81
3.3.3. Sub-clade signature deletions within the genomes of Pakistan *V. cholerae* 83
3.3.4. Diversity within *V. cholerae* circulating in Kenya 84
3.3.5. The phylogeny of Kenyan *V. cholerae* based on whole genome sequences 86
3.3.6. Genomic features of Kenyan O1 El Tor sub-clades 90
3.3.7. A novel *ctx*B gene in some Kenyan non-O1 environmental isolates 93
3.3.8. Whole genome phylogeny of Mexican strains 94
3.3.9. Genomic islands and new markers in the Mexican *V. cholerae* genomes 97
3.4. Conclusion and lessons from the regional case studies 100

4. The genetic basis of serotype variation in *V. cholerae* samples during clinical trial in Kolkata, India 103
4.1. Introduction 103
4.2. Results and discussion 105
4.2.1. *wbe*T sequence analysis 105
4.2.2. Mapping *V. cholerae* from a vaccine trial performed in Kolkata to the global El Tor phylogeny 109
4.3. Lessons learned and questions arising from this study 111

5. Expanded analysis of the seventh pandemic *V. cholerae* lineage and design of PCR based SNP typing assays 113
5.1. Introduction 113
5.2. Results and discussion 117
5.2.1. SNPs for genotyping 117
5.2.1.1. Selection of canonical SNPs 117
5.2.1.2. Phylogenetic analysis on selected SNPs 124
5.2.2. Phylogeny expansion and MLPS kits 125
5.2.2.1. Global dissemination of wave-3 in 3 sub-waves 125
5.2.2.2. Design of the MLPA based SNP-genotyping assays 130
5.3. Lessons learned from the expanded phylogeny and importance of SNP genotyping 131

6. Conclusion and future directions 134
6.1. Conclusion 134
6.2. Future directions 135
6.2.1. Further expansion of the sequenced *V. cholerae* collection 135
6.2.2. Studies investigating the evolution of *V. cholerae* within cities, countries and continents 136

6.2.3. A combined transcriptomics and proteomics study of intestinal tissues taken from mice at different stages of *V. cholerae* infection 137
6.2.4. A study designed to investigate household and community level spread of *V. cholerae* 138

Methods 139

References 142

Appendix 154

Figures

Chapter 1 figures

1.1 Timeline showing pandemics....3
1.2 One of the first maps showing the spread of cholera....4
1.3 *Vibrio cholerae* bacterium....7
1.4 O-antigen serogroups and their properties....8
1.5 *V. cholerae* life cycle....10
1.6 *V. cholerae* circulation between host and environment....12
1.7 Cholera symptoms....14
1.8 *V. cholerae* colonies on selective media....15
1.9 Molecular mechanism of working of cholera toxin....19
1.10 Expression of virulence factors at different times....20
1.11 Genetic structure of CTX phage....22
1.12 Genetic differences between different CTX phages....23
1.13 Toxin co-regulated phage gene cluster....24
1.14 Genetic structure of Vibrio pandemicity island - 2....25
1.15 Genetic structure of SXT....27
1.16 Possible arrangements of genes within CTX....28
1.17 Two chromosomes of *V. cholerae* genome....35

Chapter 2 figures

2.1 Global *V. cholerae* phylogeny....45
2.2 Phylogenetic comparison of non-conventional O1 strains....46
2.3 Phylogeny of the seventh pandemic *V. cholerae* O1 El Tor lineage....48
2.4 Linear regression plot for the seventh pandemic and its waves....49
2.5 Bayesian based phylogenetic tree for the seventh pandemic....51
2.6 Spread of the seventh pandemic plotted on the world map....53
2.7 SXT variation plot....62
2.8 Comparison of the seventh pandemic and SXT tree....63
2.9 Gene flux within the seventh pandemic....64
2.10 Insertion site of the West African South American island -1 (WASA-1)....65
2.11 Recombination in the WASA-1 cluster....67

Chapter 3 figures

3.1 Pakistan strains in the seventh pandemic phylogeny....80
3.2 Spread of cholera in Pakistan....81
3.3 Scatter plot of root-to-tip distance *vs.* date of isolation....83
3.4 Scatter plot of root-to-tip distance *vs.* river source in Pakistan....83
3.5 Sites from where Kenyan samples were collected....85
3.6 Kenyan strains in the global phylogeny....87
3.7 Phylogeny of the Kenyan clade and its sub-clades....88
3.8 Novel *ctxB* of the Kenyan non-O1 strain....93
3.9 Mexican strains in the global phylogeny....95
3.10 Mexican strains in the seventh pandemic phylogeny....96
3.11 Alignment of the novel *ctx*B in Mexican strains....98
3.12 Phylogeny of the Mexican local endemic-2 lineage....100

Chapter 4 figures

4.1 Classification of *V. cholerae* into biotypes and serotypes....104
4.2 Distribution of different mutation types in the *wbeT* gene....106
4.3 Phylogenetic tree of *wbeT* from seventh pandemic strains....108

4.4 Percentage match and mismatch between phenotypic and genotypic results 108
4.5 Distribution of Inaba and Ogawa during a linical trial study in Kolkata, India.....110
4.6 Kolkata clinical trial strains and correlation between with their temporal and serotypic correlation 111

Chapter 5 figures

5.1 Molecular basis of working of multiplex ligation-dependent probe amplification (MLPA) technology 116
5.2 Seventh pandemic phylogeny and selection of canonical SNPs.....118
5.3 Phylogeny based on the selected SNPs.....125
5.4 Phylogeny of the expanded seventh pandemic lineage.....127
5.5 Detailed structure of the wave-3 and its 3 sub-waves.....129
5.6 Spread of strains from the expanded collection of the seventh pandemic 132

Tables

Chapter 2 tables

2.1 Global *V. cholerae* collection....44
2.2 CTX types in the seventh pandemic collection....58
2.3 Strains carrying SXT and various antibiotic resistance genes....61
2.4 Genomic islands in the seventh pandemic....66

Chapter 3 tables

3.1 Pakistan *V. cholerae* collection....73
3.2 Kenyan *V. cholerae* collection....76
3.3 Mexican *V. cholerae* collection....78
3.4 Genomic islands found in Kenyan strains....92

Chapter 5 tables

5.1 Selected canonical SNPs....123
5.2 Questions that can be answered with the selected SNPs....124
5.3 The design of MLPA kits....131

Appendix

A.1 Expanded seventh pandemic *V. cholerae* collection....154

Abstract

The Origins and Evolution of *Vibrio cholerae* O1 El Tor

Cholera, like plague, is an ancient disease of great historical importance, the spread of which was originally believed to be *via* bad air or 'miasma'. In 1854, a London based physician, John Snow, first provided epidemiological evidence for the connection between contaminated drinking water and cholera and approximately 50 years later *Vibrio cholerae*, the etiological agent of cholera was identified by Robert Koch. Cholera is still common in many regions of the world despite having one of the simplest known treatment regimes: oral rehydration. In fact, the increase in the incidence of cholera since 2007 has highlighted the risk our globalized community still faces.

Currently, scientists lack a detailed understanding of how *V. cholerae* transmits and evolves, although water is recognized as a critical factor. This PhD project exploits whole genome sequence data to investigate the evolution of *V. cholerae*, focusing on serogroup O1. Phylogenetic trees based on whole genome sequencing data obtained from over 1000 *V. cholerae* representative of seventh pandemic El Tor, classical and non-O1/O139 isolates collected from across the world where cholera occurs were used to determine evolutionary patterns and relationships. In cases where detailed phenotypic information or meta-data was available, phylogeny was used alongside clinical, phenotypic and geographical information to track and understand the global and regional spread of cholera.

The genotypic basis underpinning the basis of the Ogawa to Inaba serotype change was investigated using *V. cholerae* sampled during a phase III vaccine trial undertaken in Kolkata, India and these mechanisms were defined.

Based on my mining of the phylogeny and whole genome data on which it is based, informative SNPs were selected for the basis of a simple and mobile SNP genotyping scheme. Multiplex ligation-assisted probe amplification (MLPA) was selected as the most suitable laboratory based molecular technique to detect the canonical SNPs and two kits were designed for the use of scientific and public health communities in developing countries.

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).

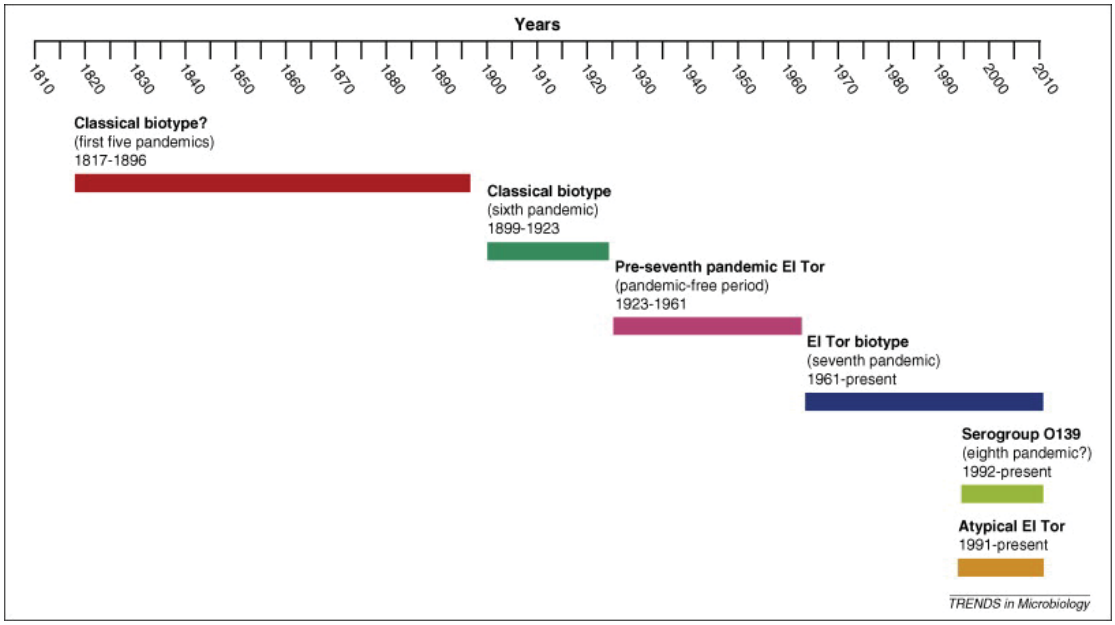


Figure 1.1: Timeline showing all the seven pandemics and hypothesized eighth pandemic or seventh sub-pandemic. Figure reproduced from (Safa, *et al.*, 2010).

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.

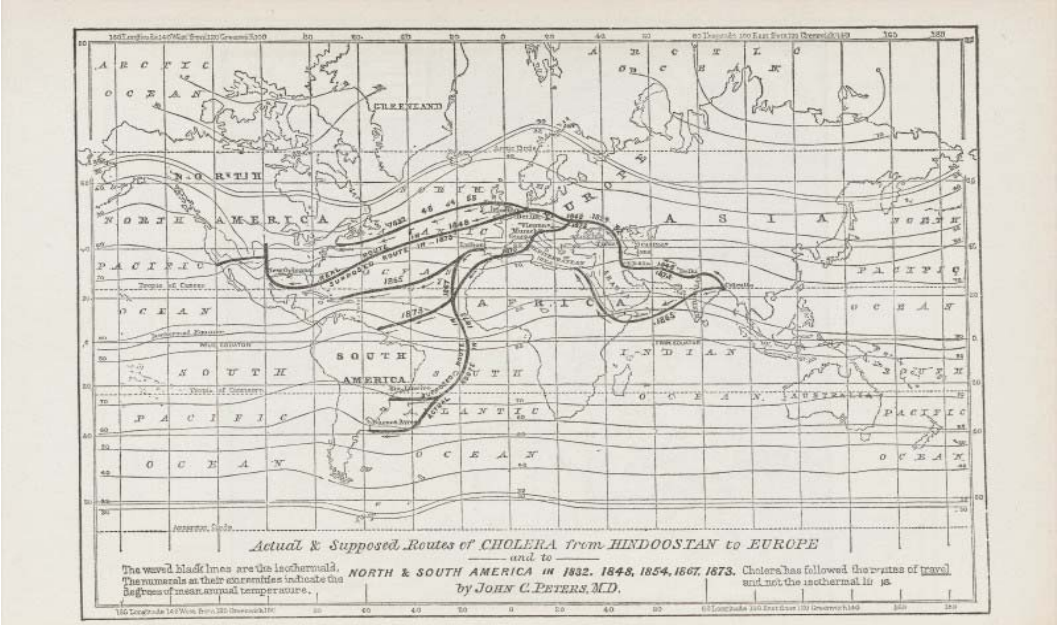


Figure 1.2: One of the first maps showing the spread of Asiatic cholera from the South Asian sub-continent to the rest of the world (Wellcome Image Library reference GC WC262 1885W47t).

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 ^{o}C but most grow well between 25-37 ^{o}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

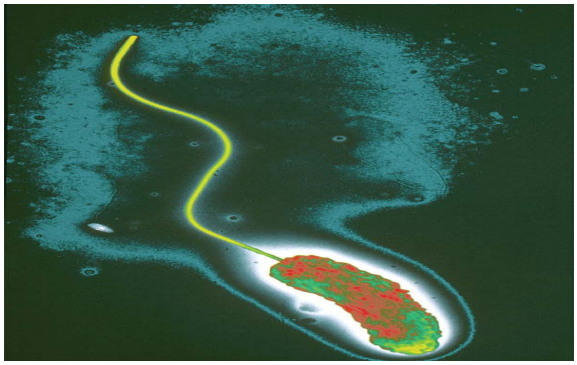


Figure 1.3: 10,000 x magnification of *V. cholerae* bacterium. Figure reproduced from Waldor *et al.* 2000 (Waldor and RayChaudhuri, 2000).

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.

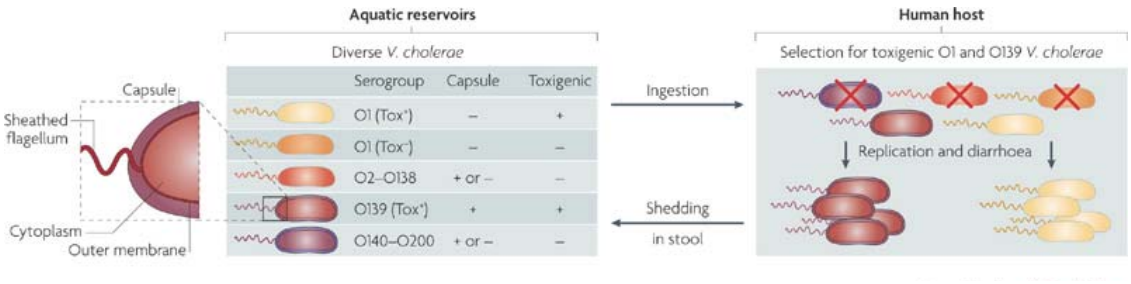


Figure 1.4: O-antigen serogroups, their main properties and the selection they may undergo when they infect humans are shown. Figure reproduced from Nelson *et al.* 2009 (Nelson, *et al.*, 2009).

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).

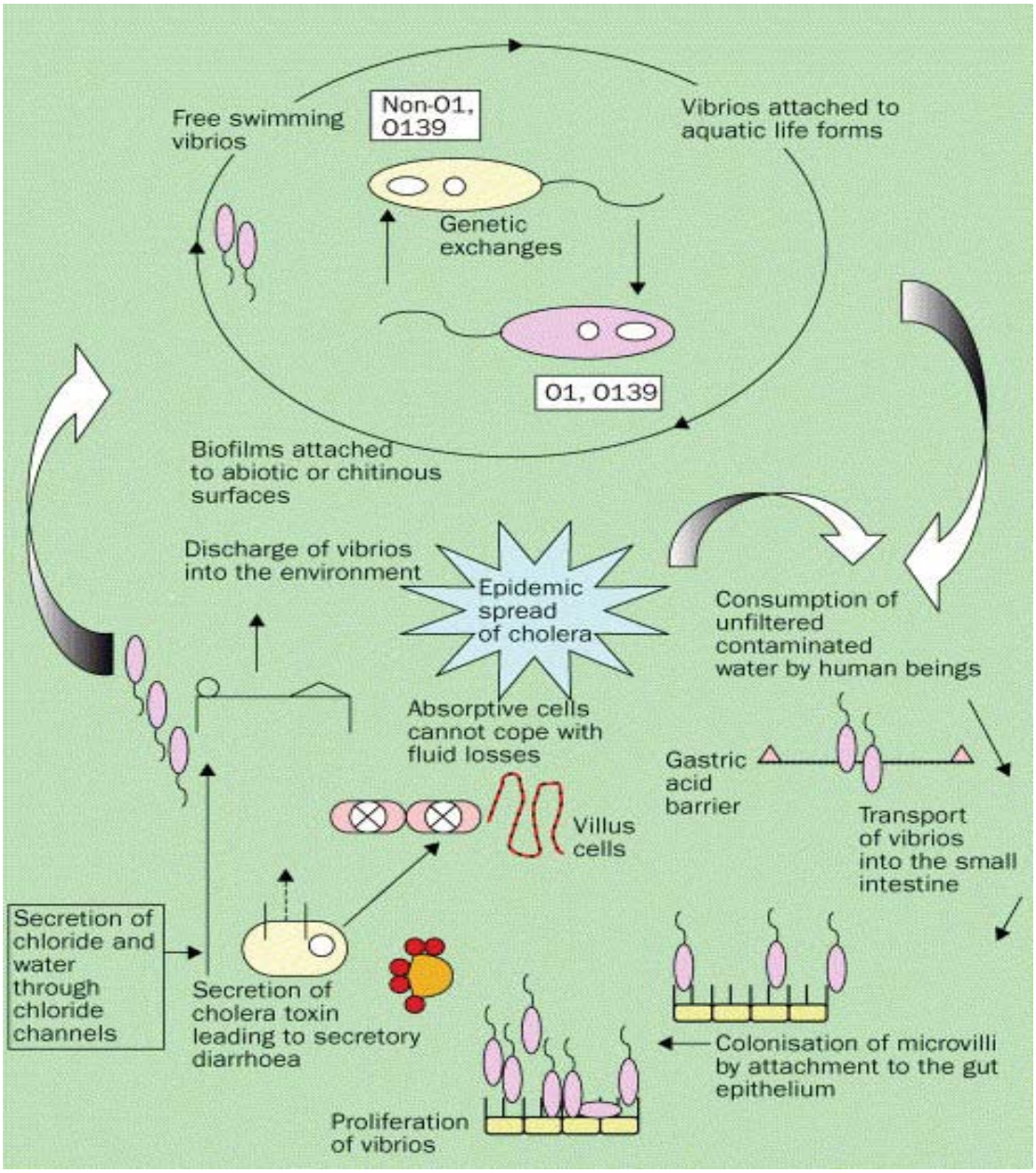


Figure 1.5: The proposed life cycle of *V. cholerae*. Figure reproduced from Sack *et al.* 2004 (Sack, *et al.*, 2004).

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.

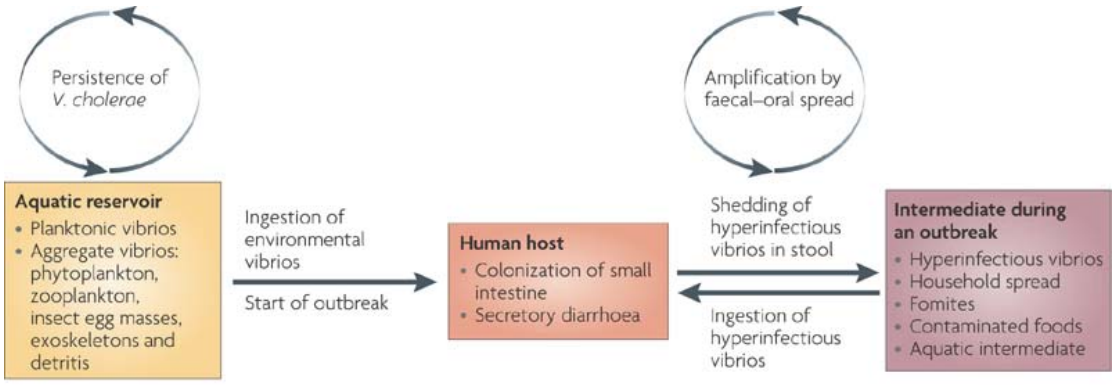


Figure 1.6: Circulation of pathogenic *V. cholerae* between environment and humans during outbreaks. Figure reproduced from Nelson *et al.* 2009 (Nelson, *et al.*, 2009).

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.

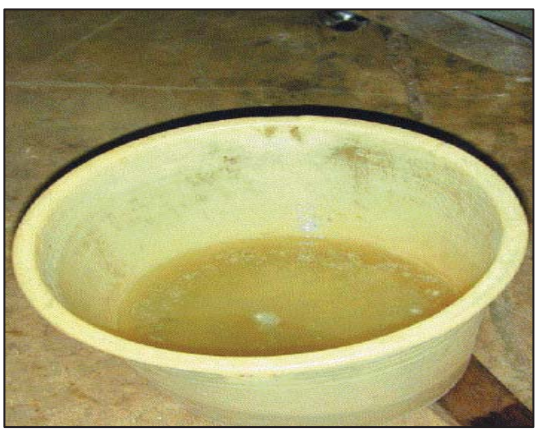


Figure 1.7: A suffering child shows typical cholera symptoms; rice water stool is collected for diagnosis and monitoring of diarrheal volume discharged. Figures reproduced from Sack *et al.* 2004 (Sack, *et al.*, 2004).

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.

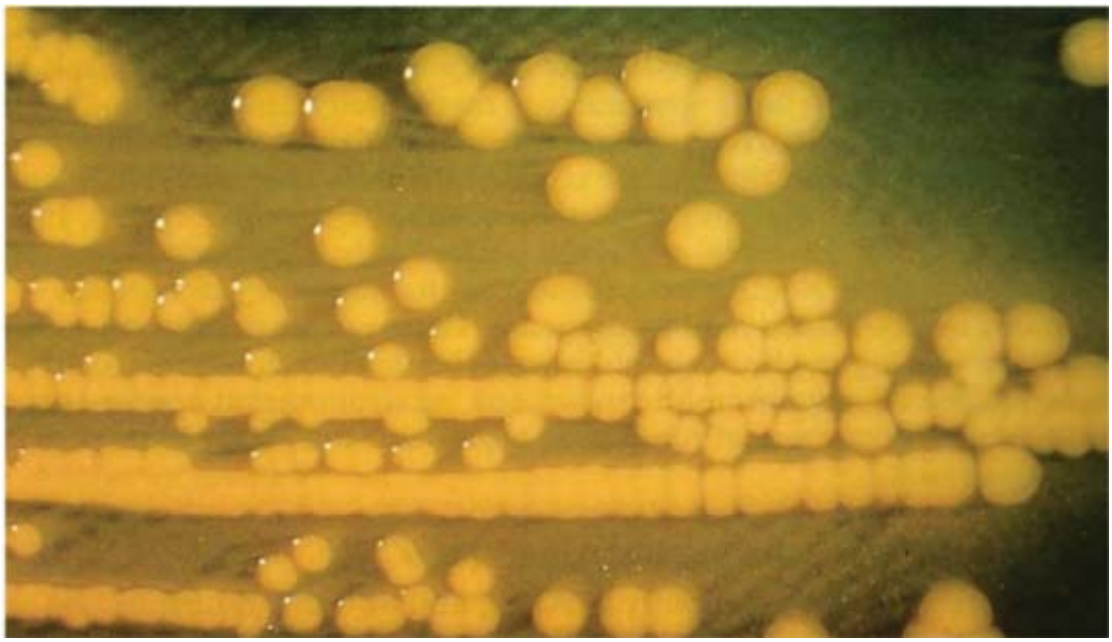


Figure 1.8: *V. cholerae* colonies on selective TCBS media. Figure sourced from Laboratory methods for the Diagnosis of *Vibrio cholerae*, Centre for Disease Control and Prevention.

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).

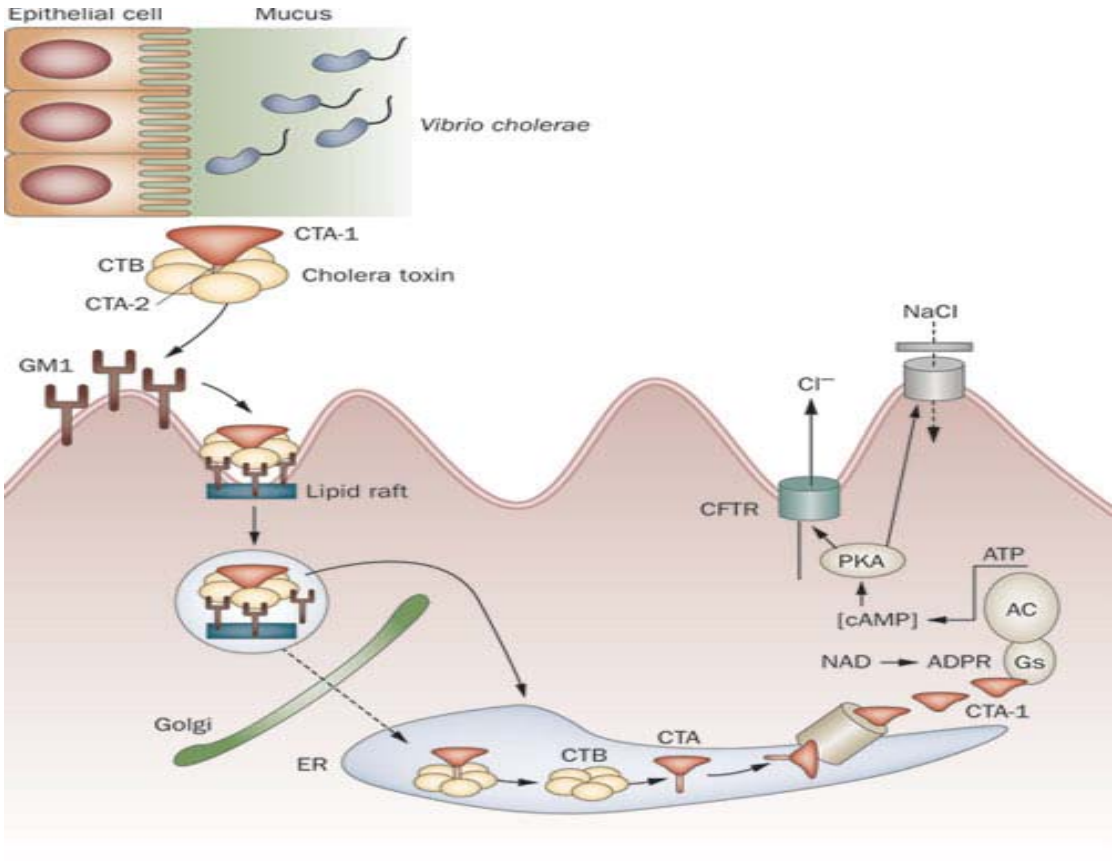


Figure 1.9: Molecular mechanism of the working of CT. Figure reproduced from Clemens *et al.* 2011 (Clemens, *et al.*, 2011).

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).

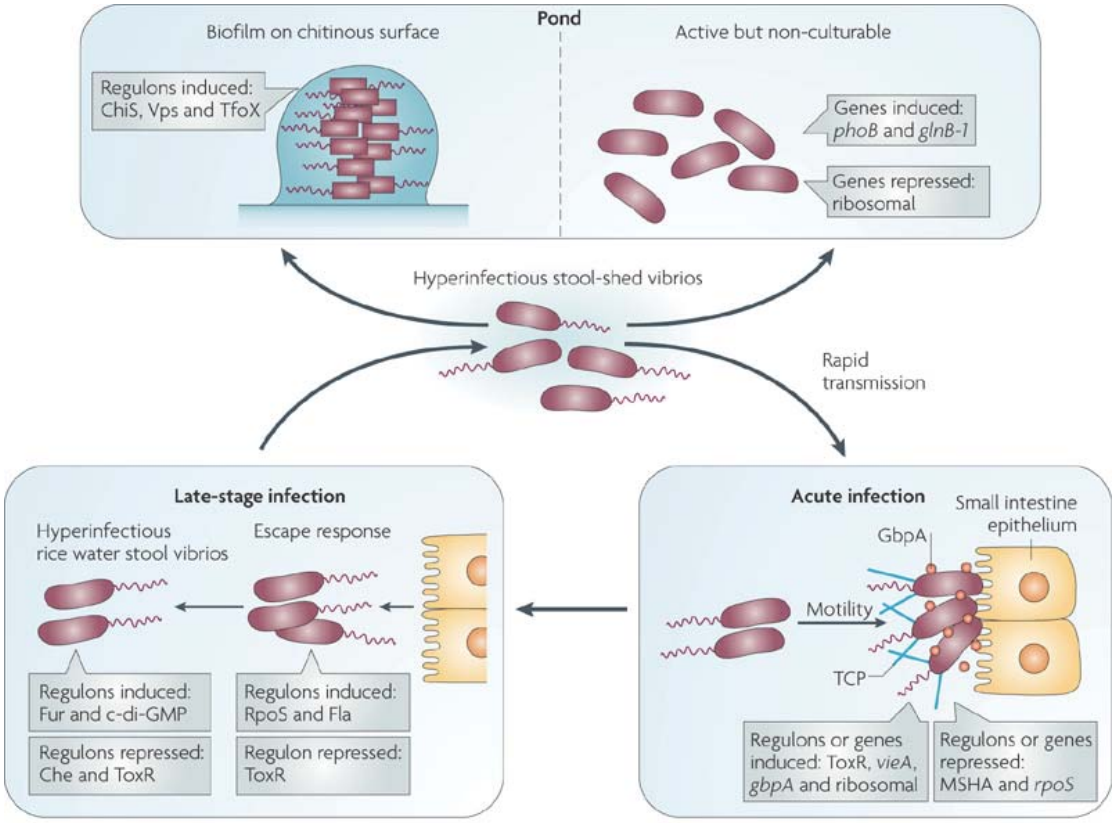


Figure 1.10: Expression of virulence factors at different times regulates the establishment of cholera infection and transmission. Figure reproduced from Nelson *et al.* 2009 (Nelson, *et al.*, 2009).

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.

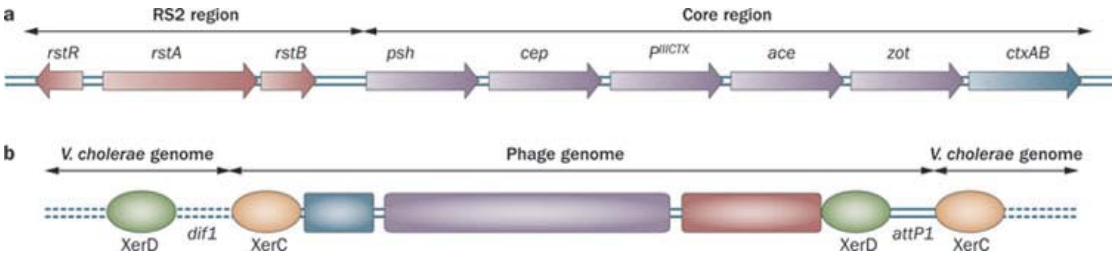


Figure 1.11: a) Arrangement of core and RS2 genes of cholera toxin phage and b) the integration sites in the *V. cholerae* chromosome. Figure reproduced from Clemens *et al.* 2011 (Clemens, *et al.*, 2011).

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).

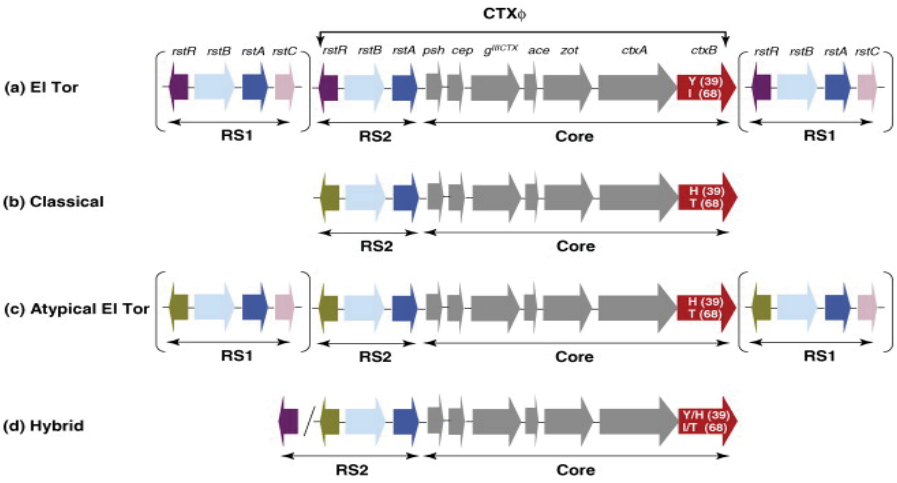


Figure 1.12: CTX phage from classical (a) and El Tor (b) strains have been reported in original respective biotypes but reports of variants of El Tor (c and d) have become more common in recent years. Figure reproduced from Safa *et al.* 2010 (Safa, *et al.*, 2010).

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.

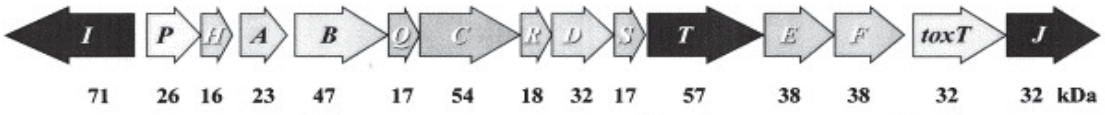


Figure 1.13: The TCP gene cluster of *V. cholerae*. Figure reproduced from Manning

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.

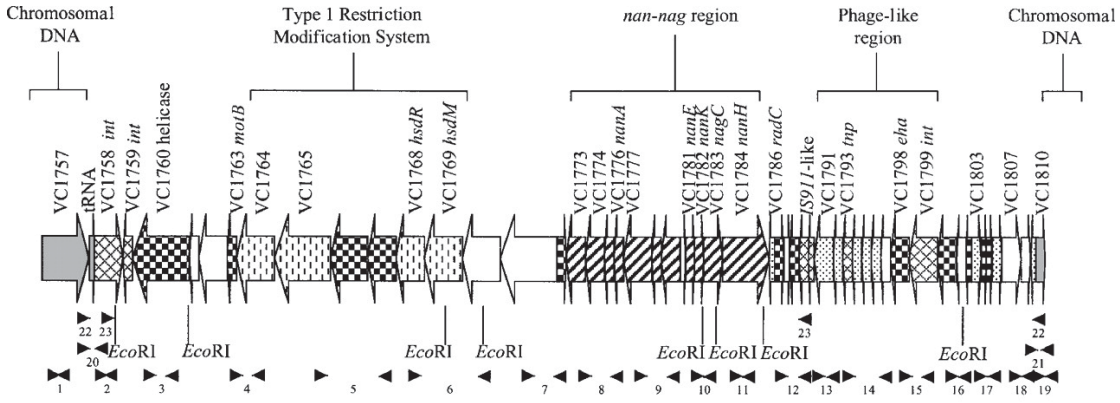


Figure 1.14: Arrangement of all the genes on the VPI-2 island, important regions are marked by their CDS number in the *V. cholerae* genome. Figure reproduced from Jermyn *et al.* 2002 (Jermyn and Boyd, 2002).

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.

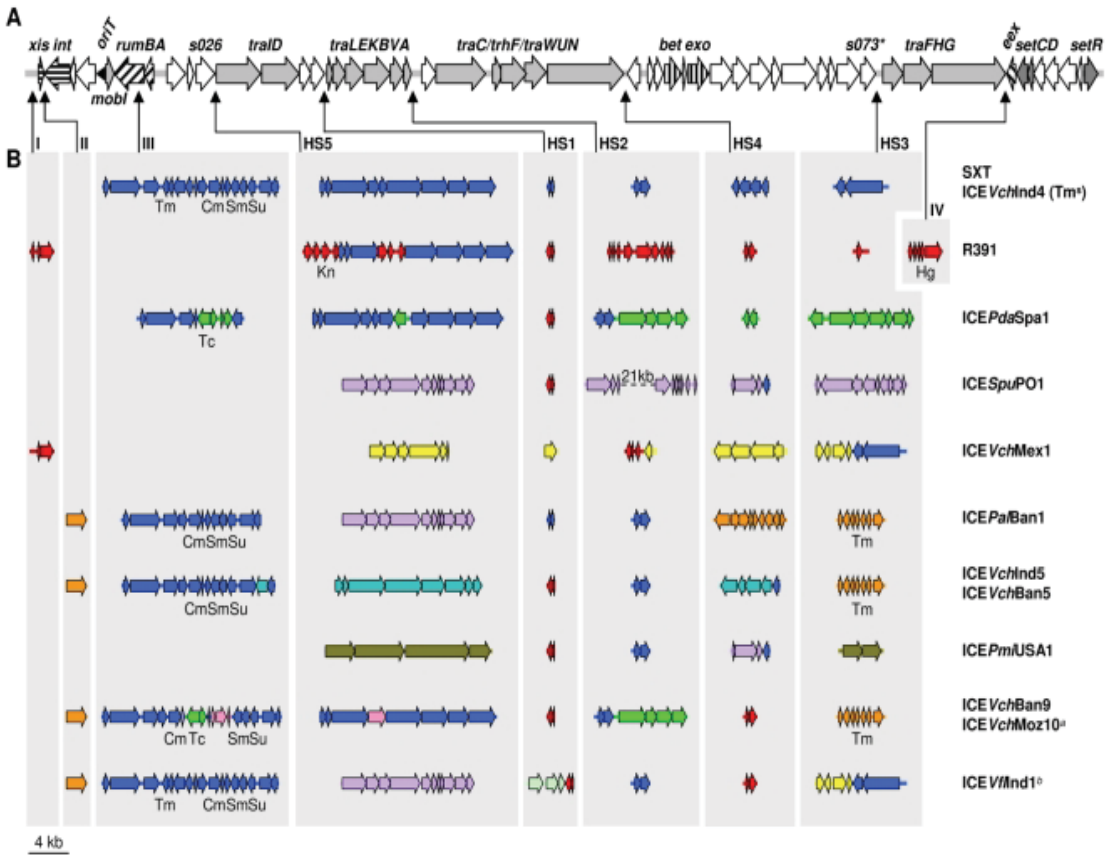


Figure 1.15: The structure of SXT/R391 family of ICE elements and the hot spot regions where different antibiotic resistance gene cassettes are inserted. Figure reproduced from Wozniak *et al.* 2009 (Wozniak, *et al.*, 2009).

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.

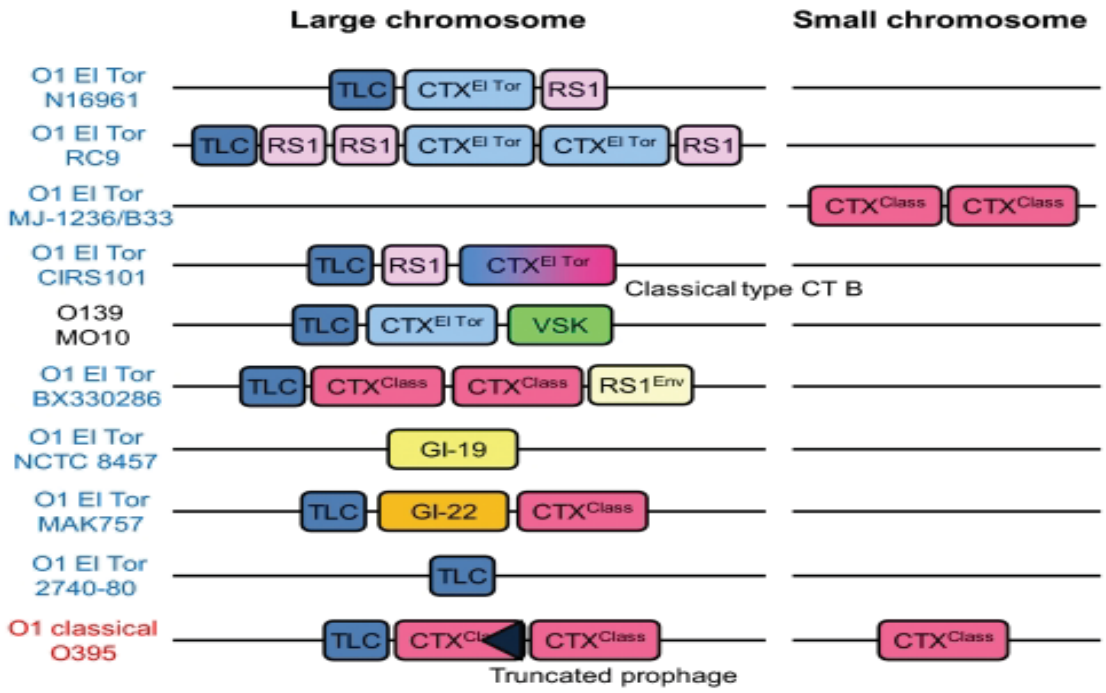


Figure 1.16: Many possible arrangements of genes within different CTX phages from different *V. cholerae* strains. Figure reproduced from Chun *et al.* 2009 (Chun, *et al.*, 2009).

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).

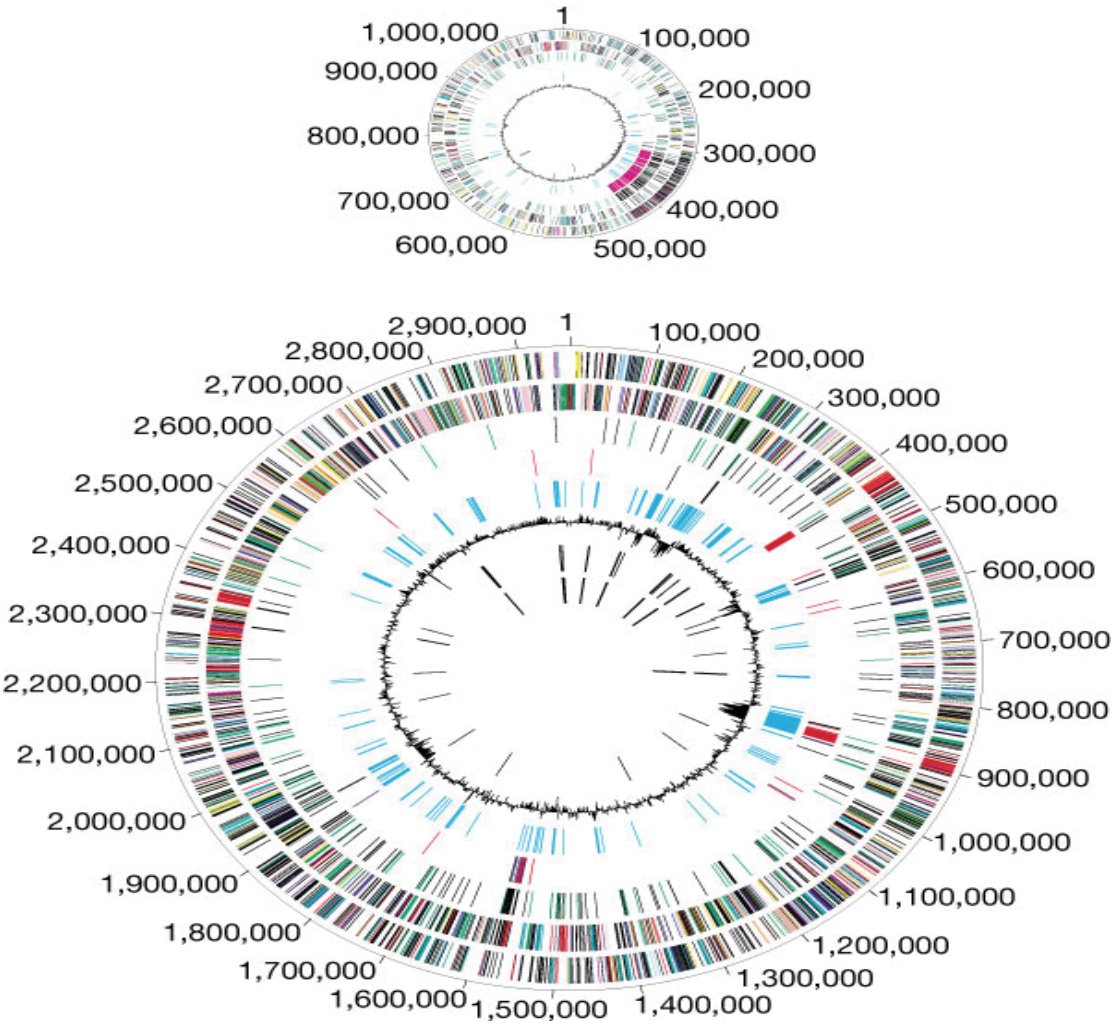


Figure 1.17: The genome of *V. cholerae* incorporates a ~1 Mb chromosome 2 (top) and ~3 Mb chromosome 1 (bottom) (Heidelberg, *et al.*, 2000). See original reference for full description of this figure.

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.

2. Genomic variation in global *V. cholerae* spanning a century

NOTE: All the isolates were collected by our global collaboration partners. The DNA was sent to the Sanger Institute for sequencing by the sequencing pipeline teams and raw short read data was made available for the analysis. The work explained in this chapter details the global phylogenetic analysis, which was done by me and therefore forms a part of my PhD thesis.

2.1 Introduction

V. cholerae is a globally important pathogen that is still endemic in many areas of the world and continues to cause cholera epidemics in others. Cholera is a severe diarrheal disease that has had a profound impact on human health for at least 1000 years (Heidelberg, *et al.*, 2000). However, since the beginning of the nineteenth century there have been reports of seven cholera pandemics, with the current (seventh) pandemic originating in 1961 from Indonesia (Lam, *et al.*, 2010; Safa, *et al.*, 2010). The latest WHO statistics (*http://www.who.int/wer*) show that 3-5 million people are affected by cholera every year with the outbreak in Haiti being recently well-publicized example (Chin, *et al.*, 2011). In Haiti, the January 2010 earthquake resulted in a break down of sanitation and hygiene systems, which gave way to the declaration of a cholera epidemic a mere 10 months later. In one month from the first report of *V. cholerae*, cholera was reported from all the states of Haiti. The most up to date figures come from a two-year surveillance study following the earthquake when Haitian public health reported 604,634 cases of infection, 329,697 hospitalizations and 7,436 deaths from cholera (Barzilay, *et al.*, 2013). The scale was such that more than 50% of recorded WHO cholera cases in 2010 and 2011 were from Haiti.

Although the species *V. cholerae* is genetically diverse, out of more than 200 O-antigen serogroups, only isolates of O1 and the recombinant derivative O139 (Chun, *et al.*, 2009; Hochhut and Waldor, 1999) can cause epidemic cholera (Chun, *et al.*, 2009). Serogroup O1 *V. cholerae* is a remarkably successful pathogen, able to infect human populations through contaminated water and food and supplies in widely diverse geographical settings. O1 strains can be further classified into two biotypes

known as classical or El Tor based on a number of biochemical and microbiological tests (see section 1.2.1 for details). It is widely accepted that the first six cholera pandemics were caused by *V. cholerae* O1 of the classical biotype but these were replaced by O1 serogroup El Tor biotype strains marking the onset of the ongoing seventh pandemic (Chin, *et al.*, 2011). Since the replacement of classical biotype strains by those of the El Tor biotype was so precipitous, many believed that the seventh pandemic strains are derived from classical strains.

Detailed epidemiology and mapping of transmission routes was compromised by a lack of informative phylogenetic markers on the *V. cholerae* genome. Traditional approaches to subtype *V. cholerae* include biochemical tests, phage typing, and low-resolution molecular typing techniques (see sections 1.2.1 and 1.2.11). CTXΦ typing has been a typing method of choice until very recently and it has led to the identification of hybrid and atypical variants of El Tor O1 where classical sequence signatures have replaced those of El Tor (Ansaruzzaman, *et al.*, 2007; Nair, *et al.*, 2002; Nair, *et al.*, 2006; Safa, *et al.*, 2010). However, recently numerous variants of CTXΦ have been described making this typing scheme unreliable (see section 1.2.11). The currently used typing techniques, including the gold standard PFGE, are based on the variable regions or mobile genetic elements and therefore it is difficult to use this information to provide a single cohesive description of the longitudinal spread and evolution of *V. cholerae*. Moreover, since the seventh pandemic strains are clonal and have considerably low genetic diversity, currently the best and the only way to accurately find the true relatedness and track the spread of this bacteria is by sequencing their whole genome and utilizing this information to construct robust family trees or phylogenies.

Previously in the study of Chun *et al.* (Chun, *et al.*, 2009), 23 strains were sequenced, including O1 and non-O1 *V. cholerae*. They showed that the strains clustered in 12 distinct lineages one of which was comprising of classical and El Tor. This study was based on highly diverse set of strains and had limited resolution within the seventh pandemic and other lineages. Therefore, we set out to define more precisely the global phylogeny of *V. cholerae* with particular focus on the strains from the current pandemic and an aim to understand the pattern of their global spread.

This chapter details the phylogeny of the lineage responsible for the current seventh pandemic. This work (collection of strains, meta-data and PCR based CTX analysis) was carried out in collaboration with our global cholera research partners. For my part I carried out all the genomic, phylogenetic and evolutionary analysis discussed here in this chapter.

Whole genome sequences from a representative sample of 154 *Vibrio cholerae* isolates spanning 100 years of cholera (1910-2010) were analysed using phylogenetics and individual lineages were analysed in detail to understand the evolution of individual important lineages. The intercontinental transmission of the seventh pandemic was tracked and the hypothesis that the seventh pandemic strains are derivatives of the previous pandemic strains, i.e. the classical biotype lineages, was put to test. Bayesian phylogenetic analysis was used to date important phylogenetic time-points and important nodes in the phylogenetic tree. The data from this study also highlighted the importance of antibiotic resistance as a driver shaping the evolution of current pandemic strains.

2.2 Bacterial isolates

Representative El Tor isolates were collected over the past four decades and compared to previously reported and novel classical and non-O1 genome sequences (Chin, *et al.*, 2011; Chun, *et al.*, 2009). Almost all of the isolates in our diverse collection were from patients with severe cholera diarrhea contracted from contaminated water or food. The exceptions being four isolates (A209, A213, A217 and A219), which originated from diarrheal cases linked to the US Gulf Coast. The isolate BX330286 included in this study was isolated from a water sample in Australia by Chun *et al.* (Chun, *et al.*, 2009) whereas all the novel sequenced isolates were of clinical origin. All isolates included in this analysis were serogroup O1, except A330 and A383, which belong to the O139 serogroup. Five isolates (A4, A49, A59, A60 and A66) had been subjected to extensive passage in the laboratory. Table 2.1 lists all the strains included in this analysis.

Strain Name	Isolation place	Isolation Year	Serotype	Original ID	Accession Number
A330	India	1993	O139	A330	ERS013124
A383	Bangladesh	2002	O139	A383	ERS013125
A488(2)	Bangladesh	2006	Ogawa	A488	ERS013129

V5	India	1989	Ogawa	V5	ERS013130
V109	India	1990	Ogawa	V109	ERS013131
V212-1	India	1991	Ogawa	V212-1	ERS013132
VC51	India	1992	Ogawa	VC51	ERS013133
MBN17	India	2004	Inaba	MBN17	ERS013134
MG116025	Bangladesh(M)	1991	Ogawa	MG116025	ERS013135
MJ1485	Bangladesh(M)	1994	Inaba	MJ1485	ERS013126
MBRN14	India	2004	Ogawa	MBRN14	ERS013127
GP8	India	1970	Inaba	GP8	ERS013128
GP16	India	1971	Inaba	GP16	ERS013136
GP60	India	1973	Ogawa	GP60	ERS013137
GP106	W.Germany	1975	Ogawa	GP106	ERS013140
GP140	Malaysia	1978	Ogawa	GP140	ERS013141
GP143	Bahrain	1978	Inaba	GP143	ERS013142
GP145	India	1979	Inaba	GP145	ERS013143
PRL5	India	1980	Ogawa	PRL5	ERS013145
GP152	India	1979	Inaba	GP152	ERS013146
IDHO1'726	India	2009	Ogawa	IDHO1'726	ERS013147
PRL18	India	1984	Ogawa	PRL18	ERS013138
PRL64	India	1992	Ogawa	PRL64	ERS013139
A46	N.I	1964	Ogawa	A46	ERS013160
A49	N.I	1962	Inaba	A49	ERS013161
A50	Bangladesh	1963	Ogawa	A50	ERS013164
A51	Egypt	1949	Ogawa	Cairo 50	ERS013165
A57	India	1980	Ogawa	U10198	ERS013166
A59	India	1970	Inaba	A59	ERS013167
A60	Thailand	1958	Inaba	A60	ERS013168
A61	India	1970	Inaba	A61	ERS013169
A66	Bangladesh	1962	Inaba	A66	ERS013170
A68	Egypt	1949	Inaba	Cairo 48	ERS013171
A70	Bangladesh	1969	Inaba	G28190	ERS013162
A76	Bangladesh	1982	Inaba	X19850	ERS013163
A103	N.I	1990	Inaba	V584	ERS013172
A109	N.I	1990	Ogawa	V588	ERS013173
A111	N.I	1990	Inaba	V591	ERS013176
A130	India	1989	Ogawa	IDH-11	ERS013177
A131	India	1989	Ogawa	IDH-12	ERS013178
A152	Mozambique	1991	Ogawa	VC1	ERS013179
A154	Mozambique	1991	Ogawa	VC3	ERS013180
A155	Mozambique	1991	Inaba	VC3 no hem	ERS013181
A177	Colombia	1992	Inaba	602	ERS013182
A180	Colombia	1992	Inaba	1388	ERS013183
A184	Colombia	1992	Ogawa	6216	ERS013174
A185	Colombia	1992	Ogawa	6216 no hem	ERS013175
A186	Argentina	1992	Ogawa	S122	ERS013184
A193	Bolivia	1992	Ogawa	S132	ERS013185
A200	Argentina	1992	Ogawa	F14	ERS013188
A201	Argentina	1992	Inaba	BsAs110	ERS013189
A209	Florida	1980	Inaba	2741-80	ERS013190
A213	Georgia	1984	Inaba	0917-84	ERS013191
A215	California	1985	Inaba	2483-85	ERS013192
A217	Louisiana	1986	Inaba	2469-86	ERS013193
A219	Georgia	1986	Inaba	2538-86	ERS013194

A231	Mexico	1991	Inaba	VC21R	ERS013195
A232	Mexico	1991	Inaba	VC22S	ERS013186
A241	Vietnam	1989	Inaba	43/89	ERS013187
A245	Vietnam	1989	Ogawa	148/89	ERS013196
A279	Sweden	1990	Inaba	K216/92	ERS013197
A316	Argentina	1993	Ogawa	SO1419	ERS013200
A325	Argentina	1993	Inaba	B1/W	ERS013201
A346(2)	Bangladesh	1994	Ogawa	A346	ERS013202
A389	Bangladesh(M)	1987	Inaba	VM11647	ERS013203
A390	Bangladesh(M)	1987	Ogawa	VM12229	ERS013204
A397	Bangladesh(M)	1987	Ogawa	VM14169	ERS013205
A481	Djibouti	2007	Inaba	1	ERS013206
A482	Djibouti	2007	Inaba	2	ERS013207
A483	Djibouti	2007	Inaba	3	ERS013198
A487(2)	Bangladesh	2007	Inaba	A487	ERS013199
4110	Vietnam	1995	Inaba	IB4110	ERS013252
4111	Vietnam	2002	Inaba	IB4111	ERS013253
4322	India	2004	Inaba	IB4322	ERS013254
4642	India	2006	Inaba	IB4642	ERS013255
4670	Bangladesh	1991	Inaba	MG116926	ERS013256
4672	Bangladesh	2000	Ogawa	E1781	ERS016137
4122	Vietnam	2007	Ogawa	IB4122	ERS013264
4605	India	2007	Ogawa	IB4605	ERS013257
4656	India	2006	Ogawa	IB4656	ERS013258
4675	Bangladesh	2001	Ogawa	E1978	ERS013259
4679	Bangladesh	1999	Ogawa	AR-32732	ERS013260
4663	Bangladesh	2001	Ogawa	MQ1273	ERS013261
4661	Bangladesh	2001	Ogawa	MQ4	ERS013263
4660	Bangladesh	1994	Ogawa	VC073	ERS013262
6180	Nairobi	2007	Inaba	6180	ERS013208
6210	Nairobi	2007	Inaba	6210	ERS013218
6201	Nairobi	2007	Inaba	6201	ERS013217
6197	Nairobi	2007	Inaba	6197	ERS013216
6196	Nairobi	2005	Inaba	6196	ERS013215
6195	Nairobi	2005	Inaba	6195	ERS013214
6194	Nairobi	2007	Inaba	6194	ERS013213
6193	Nairobi	2005	Inaba	6193	ERS013212
6215	Kakuma	2005	Inaba	6215	ERS013211
6214	Kakuma	2007	Inaba	6214	ERS013210
6191	Nairobi	2005	Inaba	6191	ERS013209
6212	Kakuma	2007	Inaba	6212	ERS013219
7682	Machakos	2009	Inaba	7682	ERS013220
7687	Machakos	2009	Inaba	7687	ERS013226
7686	Machakos	2009	Inaba	7686	ERS013225
7685	Machakos	2009	Inaba	7685	ERS013224
7684	Machakos	2009	Inaba	7684	ERS013221
1346	Mozambique	2005	Inaba	IB1346	ERS013265
4551	India	2007	Ogawa	IB4551	ERS013266
4623	India	2007	Ogawa	IB4623	ERS013267
4593	India	2007	Ogawa	IB4593	ERS013268
4538	India	2007	Inaba	IB4538	ERS013269
4339	India	2004	Ogawa	IB4339	ERS013270
4121	Vietnam	2004	Ogawa	IB4121	ERS013271

4113	Vietnam	2003	Inaba	IB4113	ERS013273
4585	India	2007	Ogawa	IB4585	ERS013232
4552	India	2007	Ogawa	IB4552	ERS013233
4488	India	2006	Ogawa	IB4488	ERS013234
4784	Tanzania	2009	Ogawa	IB4784	ERS013235
4600	India	2007	Ogawa	IB4600	ERS013236
4646	India	2007	Ogawa	IB4646	ERS013237
4662	Bangladesh	2001	Ogawa	IB4662	ERS013238
4519	India	2005	Ogawa	IB4519	ERS013239
4536	India	2007	Ogawa	IB4536	ERS013240
1362	Mozambique	2005	Ogawa	IB1362	ERS013241
1627	Mozambique	2005	Ogawa	IB1627	ERS013242
GP160	India	1980	Ogawa	GP160	ERS013243
A4	N.I	1973	Inaba	1824	ERS013244
A5	Angola	1989	Inaba	SBL	ERS013245
A6	Indonesia	1957	Inaba	C5	ERS013246
A10	Bangladesh	1979	Ogawa	T20567	ERS013247
A18	India	1977	Inaba	Phil6973	ERS013248
A19	Bangladesh	1971	Inaba	N16961	ERS013249
A22	Bangladesh	1979	Inaba	T19479	ERS013250
A27	Peru	1991	Inaba	174	ERS013251
A29	Peru	1991	Inaba	175	ERS013274
A31	Peru	1991	Inaba	176	ERS013275
A32	Peru	1991	Inaba	176 no hem	ERS013276
A346(1)	Bangladesh	1994	Ogawa	A346	ERS013278
A488(1)	Bangladesh	2006	Ogawa	A488	ERS013279
A487(1)	Bangladesh	2007	Inaba	A487	ERS013281
MG116226	Bangladesh(M)	1991	Ogawa	MG116226	ERS013282
A215	California	1985	Inaba	2483-85	ERS013277
A325	Argentina	1993	Inaba	B1/W	ERS013280
N16961	Bangladesh	1975	Inaba	N16961	AE003852/AE003853
M66	Indonesia	1937	N.I	M66	CP001233/CP001234
2010EL_1786	Haiti	2010	Ogawa	2010EL_1786	AELH00000000.1
2010EL_1792	Haiti	2010	Ogawa	2010EL_1792	AELJ00000000.1
2010EL_1798	Haiti	2010	Ogawa	2010EL_1798	AELI00000000.1
B33	Mozambique	2004	Ogawa	B33	ACHZ00000000
CIRS101	Bangladesh	2002	Inaba	CIRS101	ACVW00000000
MJ1236	Bangladesh(M)	1994	Inaba	MJ1236	CP001485/CP001486
MO10	India	1992	O139	MO10	AAKF03000000
RC9	Kenya	1985	Ogawa	RC9	ACHX00000000
BX330286	Australia	1986	Inaba	BX330286	ACIA00000000
MAK757	Celebes_Islands	1937	Ogawa	MAK757	AAUS00000000
NCTC_8457	Saudi_Arabia	1910	Inaba	NCTC_8457	AAWD01000000
V52(O37)	Sudan	1968	O37	V52(O37)	AAKJ02000000
2740_80	USGulfCoast	1980	Inaba	2740_80	AAUT01000000
O395_Combined	India	1965	Ogawa	O395_Combined	CP000626/CP000627
12129_1	Australia	1985	Inaba	12129	ACFQ00000000
TM11079-80	Brazil	1980	Ogawa	TM11079-80	ACHW00000000

N.I = No Information; (M) = Matlab

Table 2.1: Isolates analysed in this study are listed and each colour represents a separate lineage. All the data is publically accessible and European Nucleotide Archive (ENA) accession numbers are also provided.

2.3 Results and discussion

2.3.1 Global phylogeny of the *V. cholerae* species

Whole genome analysis was used to identify SNP based variation to construct accurate phylogeny and to identify regions of variation through acquisition of loss in the genomes of individual strains or lineages. Included in this analysis were 136 novel *V. cholerae* genomes sequenced as part of this study as well as 18 previously published genomes (2010; Chin, *et al.*, 2011; Chun, *et al.*, 2009).

A high resolution, maximum likelihood phylogeny based on genome wide SNPs was constructed using the methods based on Harris *et al* (Harris, *et al.*, 2010) (see methods). The sequence reads were mapped to the finished sequence of El Tor strain N16961, a seventh pandemic *V. cholerae* isolated in Bangladesh in 1975, as reference (Heidelberg, *et al.*, 2000). Of the 154 genomes analyzed in the resulting consensus tree, 8 distinct phyletic lineages (L1-L8, Figure 2.1) were identified, 6 of which (L1-L6) incorporated O1 clinical isolates whilst the other two (L7 and L8) included an environmental isolate and an O37 serogroup isolate.

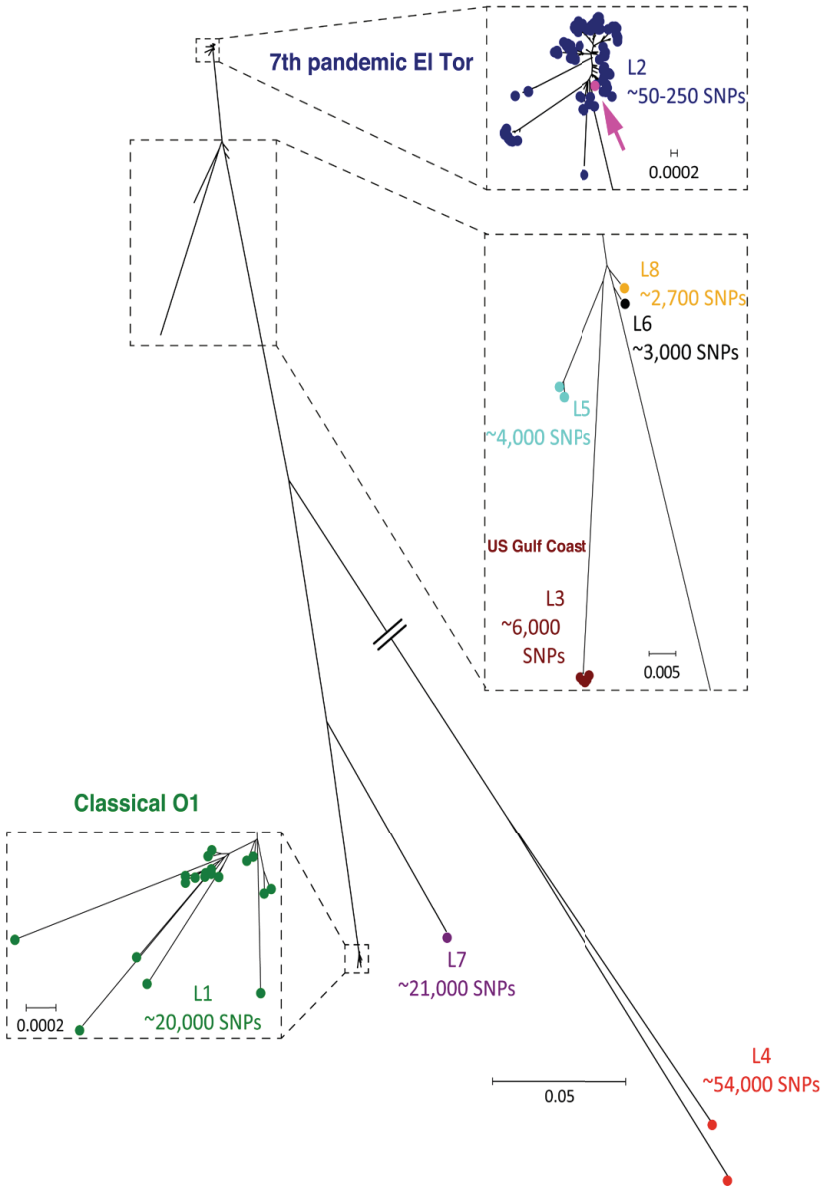


Figure 2.1: Global *V. cholerae* phylogeny of 154 isolates collected between 1910 and 2010. The maximum likelihood tree is based on SNP differences across the whole core genome and the numbers of SNP differences listed are relative to N16961 reference in L2, which is marked with an arrow. The scales are given as number of substitutions per variable site. Each of the seven lineages is shown in different colour.

Classical isolates clustered away from El Tor isolates as a distinct group termed ‘L1’. Importantly, all seventh pandemic El Tor isolates fell into a single phylogenetically distinct group named ‘L2’. The US Gulf Coast isolates clustered separately on the tree to form group ‘L3’, while the fourth group, termed ‘L4’, harbored two isolates A215

and A325 on a long branch likely to have acquired genes encoding the O1 serogroup antigen by a recombination event onto a genetically distinct genomic backbone. This lineage was similar to isolates 12129 and TM11079-80 described by Chun *et al.* ((Chun, *et al.*, 2009); Figure 2.2). While 12129 and TM11079-80 were collected from the environment, A215 and A325 were isolated from clinical samples. Chun *et al.* described their isolates as "non-conventional O1" isolates, which lack the CTX phage and signature genetic islands of pandemic strains (VPI-1 and 2, VSP-1 and 2).

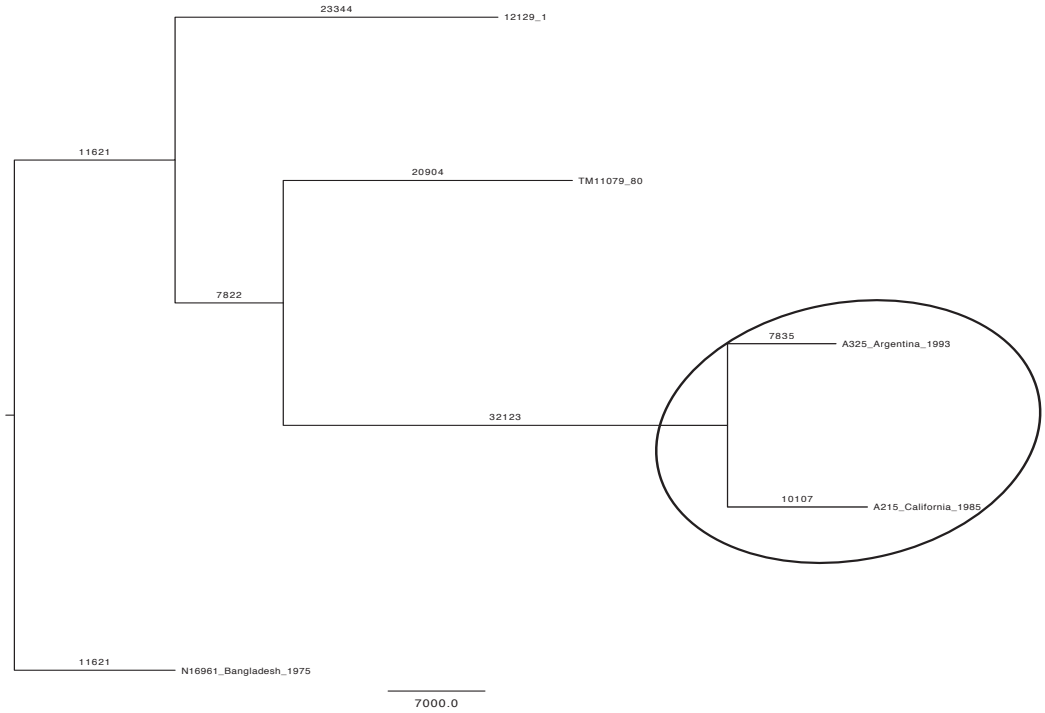


Figure 2.2: Comparison of the "non-conventional O1" strains in our collection with those from Chun *et al.* (see section 2.2). The two strains that are circled are clinical "non-conventional O1" and share common ancestor with TM11079_80. They are separated from the Chun *et al.* environmental strains 12129 and TM11079_80 by ~40,000 SNPs and form a separate lineage. The tree is rooted using N16961 reference El Tor and the scale bar indicates the number of SNPs on the branches.

Lineage L4, being a distant group with a core genome significantly different from both El Tor and classical, was used to root the phylogenetic tree (Figure 2.1). Group 'L5' constitutes M66 and MAK 757 isolated from Indonesia in 1937 that are considered pre-seventh pandemic El Tor. NCTC 8457, isolated from Sudan in 1910, was the sole representative of lineage 'L6'. V52, an O37 serogroup clinical isolate and BX330286,

a non-clinical O1 isolate, formed the two remaining lineages in the phylogenetic tree, termed ‘L7’ and ‘L8’ respectively. Isolates V52 and BX330286 were included in the study because of their interesting position in the phylogeny described by Chun *et al.* (Chun, *et al.*, 2009). V52 mapped to a location on the tree that was closer to the O1 classical lineage and BX330286 was postulated to be a hypothetical ancestor of the seventh pandemic clade (Chun, *et al.*, 2009) since it harbors genomic and pathogenicity islands found intermittently in the seventh pandemic isolates despite being of environmental origin.

From Figure 2.1 it is clear that isolates of lineage L4 were the most distantly related *V. cholerae* included in this study, differing from the reference by ~52,000 SNPs followed by L1 with ~20,000 SNP differences and L3, L5 and L6 with ~6,000, ~4,000 and ~3,000 SNP differences, respectively. V52 (L7) and BX330286 (L8) differed by ~21,000 and ~2,700 SNPs from the reference, respectively. The position of isolates on the tree and the corresponding number of SNPs clearly illustrate that groups L3, L5, L6 and L8 are more closely related to El Tor biotypes found within L2, whereas lineage L1 contains all of the classical biotypes. It is clearly evident from this analysis that the classical and El Tor clades did not originate from a recent common ancestor and instead appear to be independent derivatives with distinct phylogenetic histories.

2.3.2 Evolution of the seventh pandemic O1 El Tor *V. cholerae*

From Table 2.1 and Figure 2.1 it is clear that the L2 cluster harbored all of the 122 seventh pandemic isolates from this study, which were distinguished from each other by only 50 - 250 SNPs. The L2 cluster includes representative El Tor isolates obtained worldwide between 1957 and 2010. Consequently, with this large sample size, spanning 40 years of the seventh pandemic, a robust high-resolution phylogeny (Figure 2.3) was constructed to provide a framework for future epidemiological and phenotypic analysis of *V. cholerae* including transmission typing. The seventh pandemic phylogeny was built on the regions that were present in all the strains. Any recombination from within or outside the tree was removed in building this phylogenetic tree (see method).

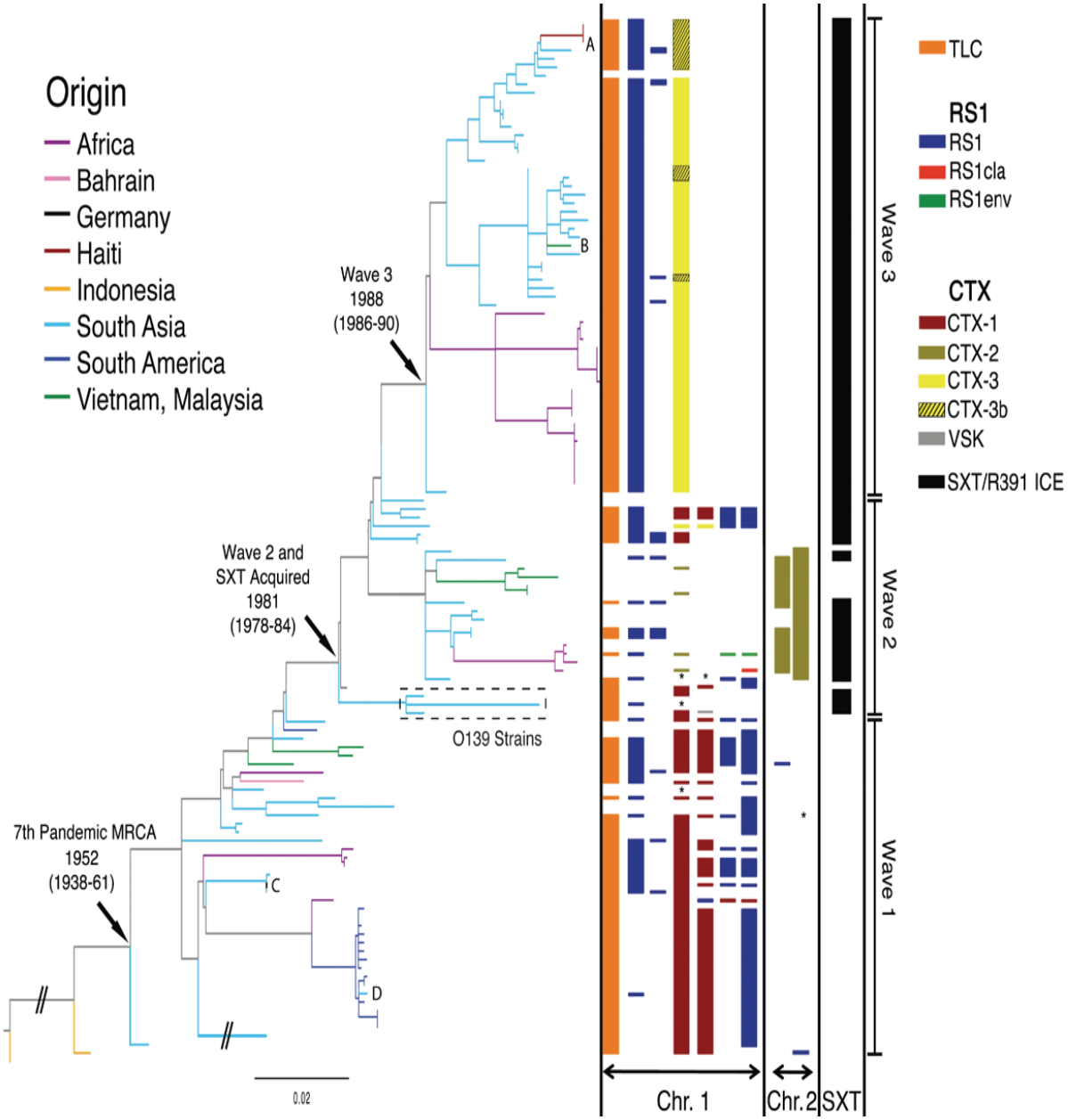


Figure 2.3: Maximum likelihood phylogeny of the seventh pandemic of L2 *V. cholerae* based on SNP differences across the whole genome, excluding likely recombination events. The tree has been rooted using M66 as an out-group and branches are coloured based on the region of isolation of the sample. CTX and SXT sequence related information is shown on the right for each strain and sporadic (or travel) transmission cases are marked as A (South Asia to Haiti), B (South Asia to Vietnam), C (South Asia to West Germany) and D (South America to South Asia). The dates of important events and nodes are derived from BEAST analysis and are the median estimates of the indicated nodes. The scale is given as number of substitutions per variable site.

Figures 2 and 3 show that the El Tor pandemic seven strains form a monophyletic lineage. When considering the dates of isolation for these strains it is clear that there is a strong temporal signature to this tree, most simply illustrated by the fact that the most divergent isolates represented in the tree are the oldest in our collection, A6 from 1957, and the most recent Haitian isolates collected by the CDC (2010) in late 2010.

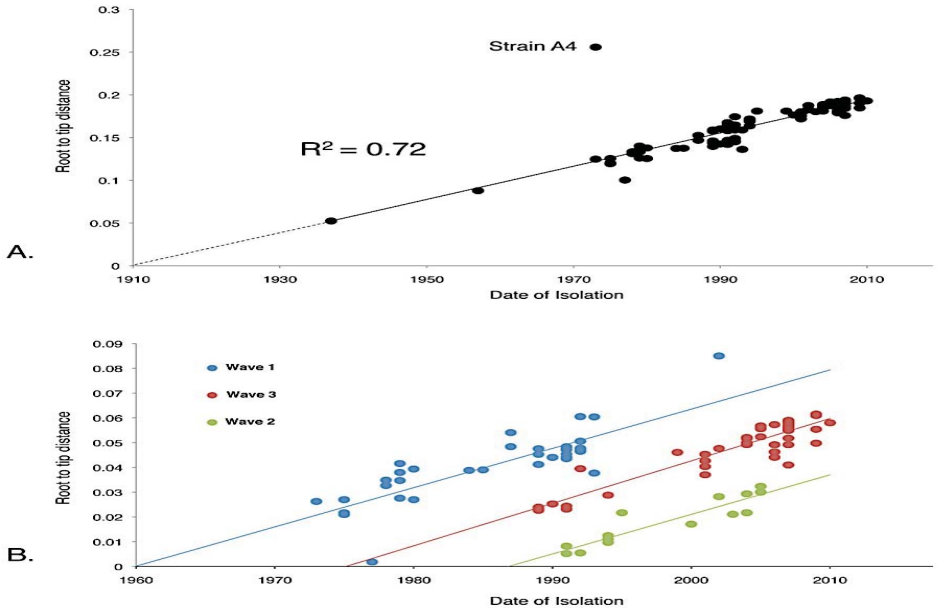


Figure 2.4: A. Root to tip distance of the seventh pandemic strains plotted against time as a linear regression plot **B.** Same analysis plot with each wave plotted separately. Isolate A4, is a 'laboratory strain' that has been multiply passaged that has been removed from the Wave 1 plot.

To accurately show that this lineage has evolved in a predictable manner a linear regression analysis was performed on all the L2 isolates. This allowed the rate of SNP accumulation to be determined based on the date of isolation and the root to tip distance (R^2 = 0.72, Figure 2.4). This analysis confirmed that *V. cholerae* has evolved in a predictable or 'clock-like' manner shown by the tight clustering of points in Figure 2.4A with an overall R^2 value of 0.7 indicating that there is a very tight correlation between the accumulation of SNPs and time. This correlation data was used to calculate that *V. cholerae* evolves at a rate of approximately 3.3 SNPs per year. The only exception to this was *V. cholerae* A4, a 'laboratory strain' isolated in 1973,

which had been subjected to repeated laboratory passage. The estimated rate of mutation for the seventh pandemic *V. cholerae* collection was 8.3 x 10^{-7} SNPs/site/year, which is about 5 and 2.5 times slower than that estimated for methicillin resistant *S. aureus* (MRSA) ST239 (Harris, *et al.*, 2010) and multi drug resistant PMEN-1 lineage of *S. pneumoniae* (Croucher, *et al.*, 2011), respectively.

Significantly, in Figure 2.3 three sub-clades of the seventh pandemic tree could be clearly seen. To formally define the structure of the tree, Bayesian Analysis for Population Structure (BAPS) (see methods) was used. BAPs analysis confirmed that within the seventh pandemic El Tor tree there were three groups, which are subsequently referred to as waves (detailed in section 2.3.3). Interestingly, when we calculated the rate of SNP accumulation independently for wave-1, wave-2 and wave-3, the rates (2.8, 2.8 and 3 SNPs/year respectively) were consistent with the rate calculated over the whole collection period (Figure 2.4B).

2.3.3 The three waves of seventh pandemic O1 El Tor *V. cholerae*

Looking at the geographical origin of the isolates detailed in Figure 2.3 it is evident that *V. cholerae* wave-1 strains were present in South Asia, South East Asia, Africa and South America between 1957 and 2002. Wave-2 and 3 strains appear geographically more restricted (reflecting the fact that *V. cholerae* epidemics since 2003 to 2010 have been restricted to South Asia, Africa and recently Haiti). What may not be clear from Figure 2.3 is that strains of wave-1 and 2 have become increasingly more rare in recent years. To test this hypothesis and to gain a dated phylogeny we performed Bayesian phylogenetic analysis of the seventh pandemic dataset using BEAST (Drummond, *et al.*, 2006). BEAST is a statistical method that uses the molecular clock information from the phylogenetic tree and superimposes the metadata like the dates of isolation and geographical information to predict the time and place of existence of the ancestral nodes. This tool was used to predict the dates of ancestral nodes at 95% confidence interval levels and the information was used to re-draw the phylogenetic tree on a time scale. It dated the most recent common ancestor responsible for the seventh pandemic to between 1827-1935 (Figure 2.5). This estimate was consistent with the predicted date of origin from the linear regression plot (1910, Figure 2.4). This also corresponds with the first El Tor biotype

strain being isolated in 1905 (Cvjetanovic and Barua, 1972). While the date of the most recent common ancestor (tMRCA) of wave-1 fell between 1938 and 1961, the MRCAs of wave-2 and wave-3 were 1978-1984 and 1986-1990 respectively (Figure 2.5).

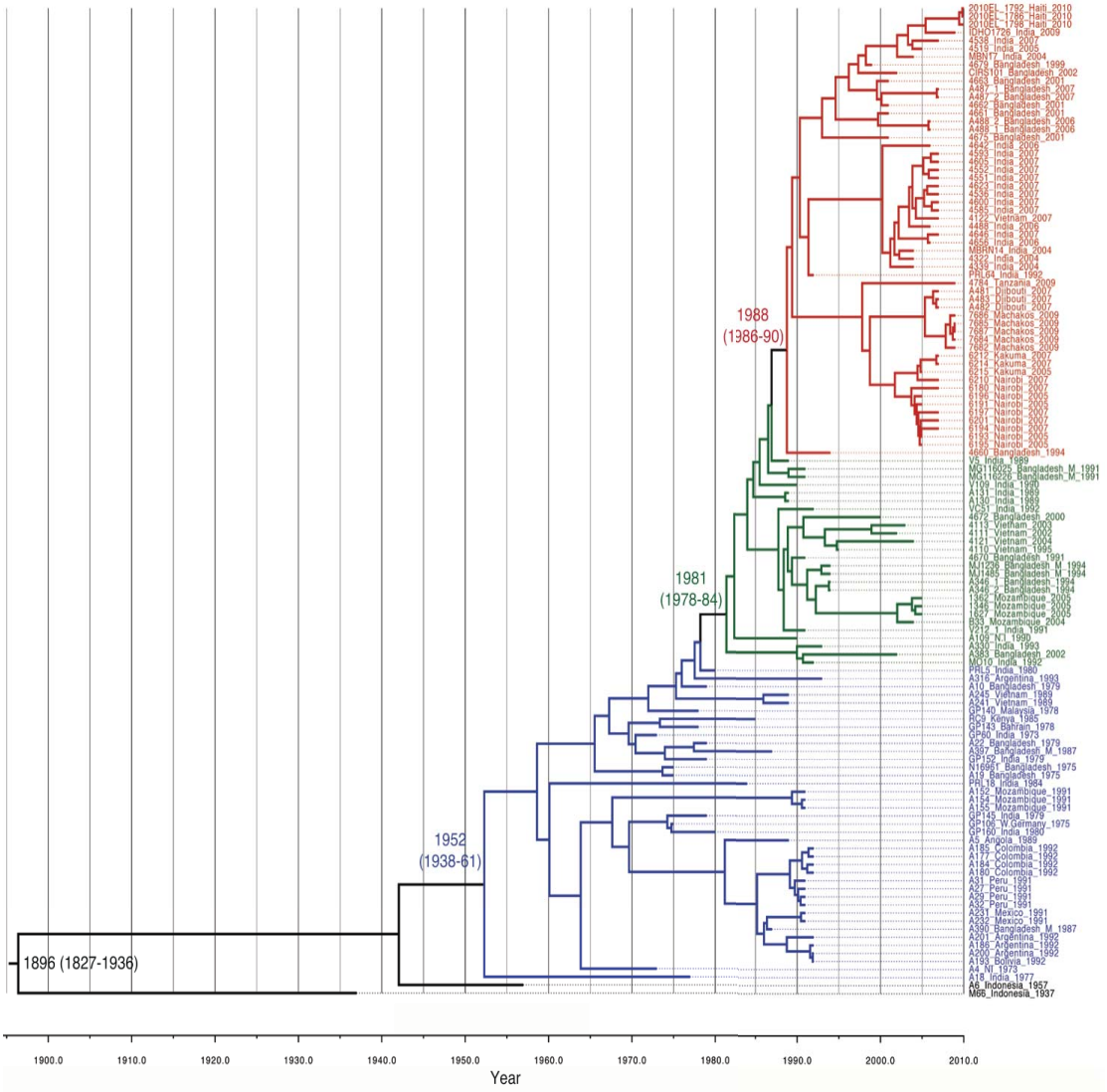


Figure 2.5: BEAST generated maximum clade credibility tree of the seventh cholera pandemic. Dates highlighted are the median values predicted with the date range representing the 95% confidence interval on the estimate. The dates shown are for the MRCAs of the El Tor lineage, wave-1, wave-2/SXT acquisition and wave-3. Blue, green and red colours represent wave-1, 2 and 3 respectively.

When considering the dated phylogeny and the geographic locations of the different isolates such as Vietnam or South Asia (Figure 2.5), it is also noticeable that wave-1 isolates were then largely replaced by either wave-3 or wave-2 clades, a phenomenon supported by previous clinical observations and phage analysis (Safa, *et al.*, 2010). The strains of wave-1 in this study originate from the beginning of the seventh pandemic until 1993 and interestingly, no strain reported after the mid 1990s clustered within the wave-1. This suggests that strains of a particular genotype (BAPS group 1 in this case) that are successful in causing outbreaks tend to disappear thus marking the end of the respective wave. Further, the phylogeny showed that eventually the strains that go extinct are replaced by strains of another SNP genotype, which co-exist alongside the strains of previous wave for a limited period. This pattern was noticed, as strains of BAPS group 2 (wave-2) and 3 (wave-3) in our study were isolated between 1989-2005 and 1994-2010 respectively. Similar to the wave-1 strain replacement, wave-2 strains were also completely replaced by wave-3 strains after their co-existence for some time, therefore, marking the end of wave-2. This overlapping but independent replacement of strains of one SNP genotype by another in the form of waves is peculiar because each time the ancestor to the new wave of strains radiates directly from the backbone of the phylogenetic tree instead of extending from the previous wave's ancestors. Moreover, the basal strain (s) in all three cases was from the South-Asian sub-continent, which strongly indicates that there is a single source from where the strains travel to the non-endemic areas, cause epidemics and then disappear to give way to a new set of strains from the same source.

To show this more clearly the distribution of the strains from the three waves were plotted onto a world map. Alongside the phylogenetic structure data (Figure 2.3), the ancestral date information from the BEAST analysis (Figure 2.5) was used to mark the closest approximate date of travel of strains from one geographical location to the other. The resulting pattern confirmed that the *V. cholerae* seventh pandemic is sourced from a single geographical location but has spread in overlapping waves (Figure 2.6). The coupling of genomic variation data and epidemiology i.e. the findings of this study tie in closely with the traditionally believed fact that the Bay of Bengal could be the seeding source of the seventh pandemic.

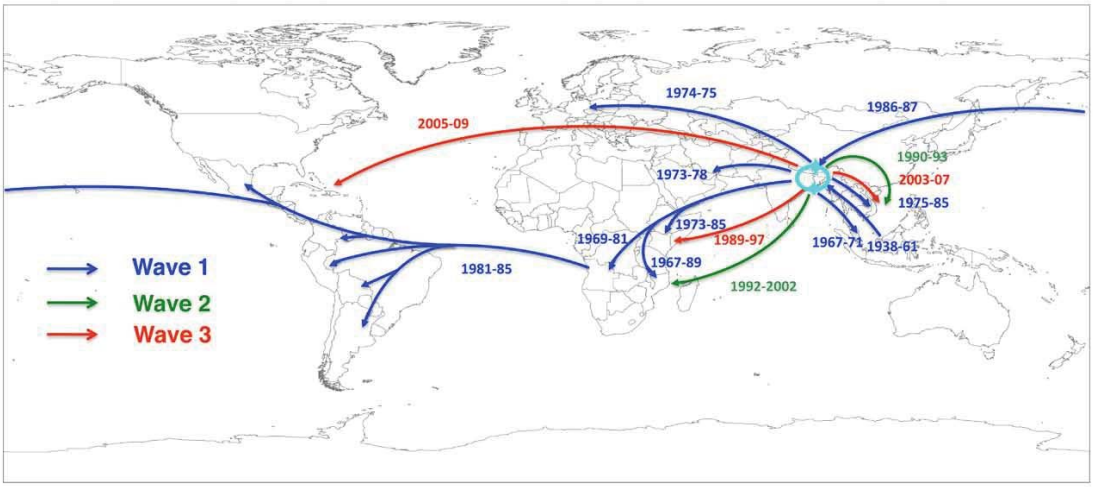


Figure 2.6: seventh Pandemic phylogenetic tree drawn on a global map to show the transmission events. The date ranges shown are derived from Bayesian analysis and represent the median values for the most recent common ancestors of the transmitted strains (later bound), and the MRCA of the transmitted strains and their closest relative from the source location (earlier bound).

2.3.4 The origins of O139 serogroup strains

As mentioned earlier (section 1.2), of the more than 200 O-antigen based serogroups of *V. cholerae* only O1 and O139 strains are known to have capacity to cause major outbreaks. Infections due to O139 serogroup strains, first reported in India and Bangladesh in 1992, surpassed the number of clinical cholera cases due to O1 infection. Many at the time saw the advent of O139 strains as the beginning of the eighth pandemic of cholera. However, by early 2000s the O139 strains largely disappeared due to yet unknown reasons. In this study we sequenced two novel O139 strains and included the sequence of previously published O139 strain MO10 (isolated in India in 1992) in our analysis to identify the origins of the O139 lineage. When mapped to the reference El Tor strain N16961 (Figure 2.3), all the O139 strains clustered within the wave-2 of L2 lineage and shared the most recent common ancestor (tMRCA) with a South Asian isolate. This analysis also confirmed the previous findings that the isolates of serogroup O139 have arisen from a homologous replacement event of their O-antigen determinant into an El Tor genomic backbone (Chun, *et al.*, 2009; Hochhut and Waldor, 1999; Lam, *et al.*, 2010). Thus, it would not

be wrong to say that O139 may represent another distinct, but spatially restricted, wave from the same common source but the lineage is clearly a derivative of *V. cholerae* O1 El Tor strains.

2.3.5 Evidence within the global phylogeny of intercontinental transmission

Other than the on going spread from the South-Asian sub-continent to the rest of the world, which was evident in the phylogeny (Figure 2.3 and 2.8), there were more examples of intercontinental transmission throughout the structure of the tree. The South American isolates formed a discrete cluster in wave-1, which also included an Angolan isolate collected in 1989. This strain was on a branch basal to the South American cluster. It is within this time period that the seventh pandemic cholera occurred in South America (Heidelberg, *et al.*, 2000), which suggests that cholera in South America could have entered from West Africa (detailed in section 2.3.9). Four incidences of sporadic transmission or traveller transmission were also clearly identifiable in the phylogenetic tree (A – South Asia to Haiti; B – South Asia to Vietnam; C – South Asia to Germany and D – South America to South Asia in Figure 2.3), indicating that non-symptomatic travellers can carry O1 El Tor *V. cholerae* and pass through regional boundaries unnoticed.

2.3.6 Patterns of gene acquisition and loss in the seventh pandemic

Previous sub-genomic sequence-based studies have focused on novel genomic islands in *V. cholerae* that are generally mobile and/or relatively unstable (Lam, *et al.*, 2010). For the first time, by virtue of this study, the *V. cholerae* SNP based phylogeny provides a robust backbone on which temporal acquisition and loss of such mobile elements could be placed and key insertion/deletions, recombination events, the variations reported in CTX, the cholera toxin operon (see section 1.2.11), the acquisition of the multi-drug resistant cassette SXT (see section 1.2.10) could be monitored.

2.3.7 Variations in CTX and their phylogenetic distribution

Sequences differentiating at least three CTX types have been previously published (Safa, *et al.*, 2010) but there is a great deal of uncertainty about how to name new CTX-types when they are discovered. To relate the distribution of CTX types with the strains across our global phylogeny it was first important to study the CTX structures of respective waves and rationally name the different CTX types. Therefore a novel scheme (as described below) was designed. In the scheme, a mutation or single base pair change in any of the CTX genes is called a new CTX type. This new expandable nomenclature, the reasoning and the scheme itself is described below:

Since the seventh cholera pandemic strains were clearly distinguished by three waves, distinct differences in their CTX genes were identified and an expandable naming system was proposed. Any new seventh pandemic *Vibrio cholerae* strain could be named using this novel, simple and expandable nomenclature scheme. The canonical El Tor CTX was called CTX-1 and the rationale below was followed to expand on this:

1) For CTX-1 to CTX-2, as there was a shift of *rst*R$^{\text{El Tor}}$ to *rst*R$^{\text{Classical}}$, *rst*A $^{\text{El Tor}}$ to *rst*A$^{\text{Classical+El Tor}}$ and *ctx*B$^{\text{El Tor}}$ to *ctx*B$^{\text{Classical}}$, it was called CTX-2.

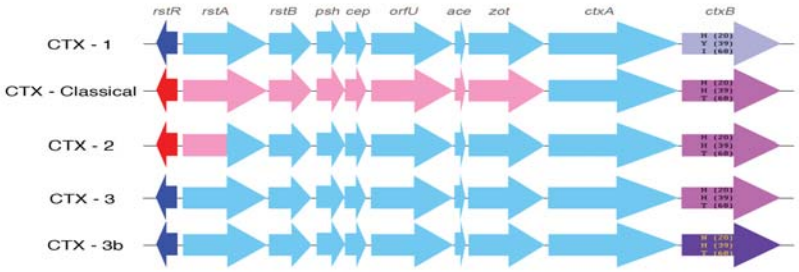
2) For CTX-1 to CTX-3, as there was a shift of *ctx*B$^{\text{El Tor}}$ to *ctx*B$^{\text{Classical}}$, it was called CTX-3.

3) For CTX-3 to CTX-3b, as there was only one SNP mutation in *ctx*B$^{\text{Classical}}$ from CTX-2 and rest was identical, it was treated as the next variant of CTX-3 and called CTX-3b.

Therefore, under this scheme, if there is a shift of any gene from one biotype to another, the new CTX will be called CTX-'n' and so will be the strains e.g. the next strains fitting this criteria will be called CTX-4. However, if there is a mutation(s) in the gene that does not lead to a shift of the gene to another biotype gene, CTX-1b,

CTX-1c or CTX-2b, CTX-2c or CTX-3b, CTX-3c and so on should be followed as appropriate.

Wave-1 isolates mostly harboured CTX-1 type. Whereas wave-2 isolates harboured CTX-2 representing a discrete cluster that show a complex pattern of accessory elements within the CTX locus (Figure 2.3) and a wide phylogeographic distribution. To date no new CTX-2 *V. cholerae* isolates have been reported since 2006 from either endemic or epidemic areas. In contrast, wave-3 isolates carrying CTX-3 or CTX-3b are the most prevalent strains today, routinely isolated from clinics treating cholera patients in all cholera reporting regions of the world. CTX types different from CTX-1, CTX-2 and CTX-3 have been reported for the O139 serogroup (Basu, *et al.*, 2000; Faruque, *et al.*, 2003; Faruque, *et al.*, 2000; Nair, *et al.*, 1994) but O139 strains are not found anymore even in the most endemic regions bordering the Bay of Bengal. The CTX types of all the seventh pandemic strains are illustrated in Figure 2.3 and listed in Table 2.2.



Wave - 1

Wave - 2

Wave - 3

Pre-7th pandemic

Strain Name	Chr 1 CTX structure	Chr 2 CTX structure	Note
2010EL_1792_Haiti_2010	TLC-RS1-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
2010EL_1786_Haiti_2010	TLC-RS1-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
2010EL_1798_Haiti_2010	TLC-RS1-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
IDH01726__India_2009	TLC-RS1-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
4538__India_2007	TLC-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
4519__India_2005	TLC-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
MBN17__India_2004	TLC-RS1-CTX-3	Blank	
4679_Bangladesh_1999	TLC-RS1-CTX-3	Blank	
CIRS101_Bangladesh_2002	TLC-RS1-CTX-3	Blank	
4663_Bangladesh_2001	TLC-RS1-CTX-3	Blank	
A487_1__Bangladesh_2007	TLC-RS1-CTX-3	Blank	
A487_2__Bangladesh_2007	TLC-RS1-CTX-3	Blank	
4662_Bangladesh_2001	TLC-RS1-CTX-3	Blank	
4661_Bangladesh_2001	TLC-RS1-CTX-3	Blank	
A488_2__Bangladesh_2006	TLC-RS1-CTX-3	Blank	
A488_1__Bangladesh_2006	TLC-RS1-CTX-3	Blank	
4675_Bangladesh_2001	TLC-RS1-CTX-3	Blank	
4642_India_2006	TLC-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
4593__India_2007	TLC-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
4605_India_2007	TLC-RS1-CTX-3	Blank	
4552__India_2007	TLC-RS1-CTX-3	Blank	
4551__India_2007	TLC-RS1-CTX-3	Blank	
4623__India_2007	TLC-RS1-CTX-3	Blank	
4536__India_2007	TLC-RS1-CTX-3	Blank	
4600__India_2007	TLC-RS1-CTX-3	Blank	
4585__India_2007	TLC-RS1-CTX-3	Blank	
4122_Vietnam_2007	TLC-RS1-CTX-3	Blank	
4488__India_2006	TLC-RS1-CTX-3	Blank	
4646__India_2007	TLC-RS1-CTX-3	Blank	
4656_India_2006	TLC-RS1-CTX-3	Blank	CTX-3b (A instead of C at nucleotide base position 58 in ctxB)
MBRN14__India_2004	TLC-RS1-RS1-CTX-3	Blank	
4322_India_2004	TLC-RS1-CTX-3	Blank	
4339__India_2004	TLC-RS1-CTX-3	Blank	
PRL64__India_1992	TLC-RS1-RS1-CTX-3	Blank	
4784_Tanzania_2009	TLC-RS1-CTX-3	Blank	
A481_Djibouti_2007	TLC-RS1-CTX-3	Blank	
A483_Djibouti_2007	TLC-RS1-CTX-3	Blank	
A482_Djibouti_2007	TLC-RS1-CTX-3	Blank	
7686_Machakos_2009	TLC-RS1-CTX-3	Blank	
7685_Machakos_2009	TLC-RS1-CTX-3	Blank	
7687_Machakos_2009	TLC-RS1-CTX-3	Blank	
7684_Machakos_2009	TLC-RS1-CTX-3	Blank	
7682_Machakos_2009	TLC-RS1-CTX-3	Blank	
6212_Kakuma_2007	TLC-RS1-CTX-3	Blank	
6214_Kakuma_2007	TLC-RS1-CTX-3	Blank	
6215_Kakuma_2005	TLC-RS1-CTX-3	Blank	
6210_Nairobi_2007	TLC-RS1-CTX-3	Blank	
6180_Nairobi_2007	TLC-RS1-CTX-3	Blank	
6196_Nairobi_2005	TLC-RS1-CTX-3	Blank	
6191_Nairobi_2005	TLC-RS1-CTX-3	Blank	
6197_Nairobi_2007	TLC-RS1-CTX-3	Blank	
6201_Nairobi_2007	TLC-RS1-CTX-3	Blank	
6194_Nairobi_2007	TLC-RS1-CTX-3	Blank	
6193_Nairobi_2005	TLC-RS1-CTX-3	Blank	
6195_Nairobi_2005	TLC-RS1-CTX-3	Blank	
4660_Bangladesh_1994	TLC-RS1-CTX-3	Blank	
V5__India_1989	Blank	Blank	
MG116025_Bangladesh_M__1991	TLC-RS1-CTX-1-CTX-1-RS1-RS1	Blank	
MG116226_Bangladesh_M__1991	TLC-RS1-CTX-1-CTX-1-RS1-RS1	Blank	
V109__India_1990	TLC-RS1-CTX-3-CTX-3-RS1-RS1	Blank	
A131_India_1989	TLC-RS1-RS1-CTX-1	Blank	
A130_India_1989	TLC-RS1-RS1-CTX-1	Blank	
VC51__India_1992	Blank	CTX-2	20bp deletion in rst A
4672_Bangladesh_2000	RS1-RS1	CTX-2-CTX-2	
4113__Vietnam_2003	CTX-2	CTX-2-CTX-2	
4111_Vietnam_2002	Blank	CTX-2-CTX-2	
4121__Vietnam_2004	Blank	CTX-2-CTX-2	
4110_Vietnam_1995	CTX-2	CTX-2-CTX-2	
4670_Bangladesh_1991	TLC-RS1-RS1	CTX-2-CTX-2	
MJ1236_Bangladesh_M__1994	Blank	CTX-2-CTX-2	
MJ1485_Bangladesh_M__1994	Blank	CTX-2	

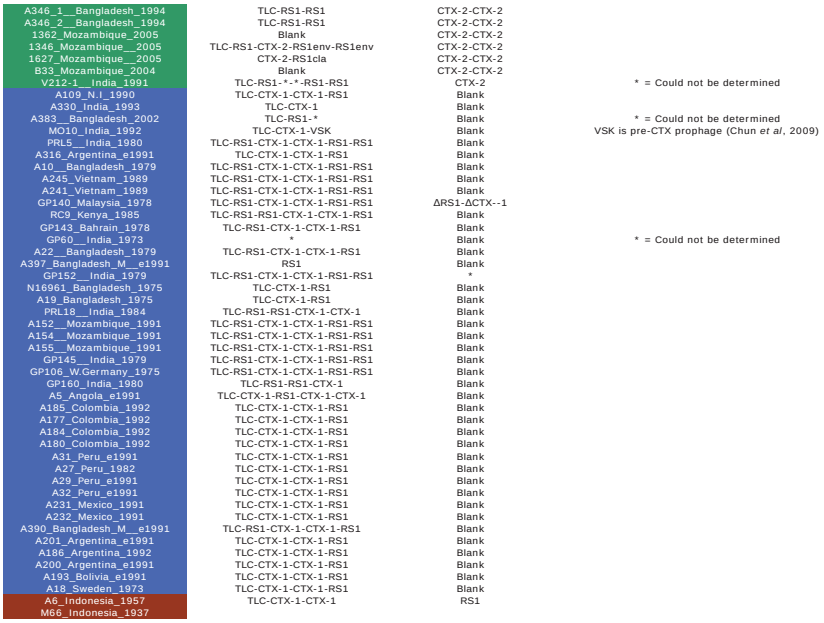


Table 2.2: Structures of CTX types (top panel) and molecular CTX type information of each isolate according to the new nomenclature scheme. In blue, green and red are the wave-1, 2 and 3 seventh pandemic strains with wave and CTX information respectively.

2.3.8 Variations in SXT and its phylogenetic distribution

SXT has played a major role in driving the spread of multiple antibiotic resistant *V. cholerae* and was consequently analysed in detail. All the strains in our collection were manually checked for the presence and absence of this ICE element insertion in the *prfC* 3 gene, the normal site specific for the insertion of SXT in the *V. cholerae* genome. When this information was superimposed onto the phylogenetic tree, the most likely first point of entry of SXT into the O1 El Tor *V. cholerae* genomic backbone could be established (Figure 2.3). Moreover, the data from the BEAST analysis showed that the date of tMRCA of wave-2 was the same as the acquisition date of the SXT element, which would have first come in between 1978 and 1984. This analysis suggested that the SXT element, which was first detected in O139 strains in 1992 and was thought to have originated within O139 strains (Hochhut and Waldor, 1999), was present in El Tor O1 ancestors at least 10 years prior to its discovery in O139.

The diversity of SXT ICE elements present in the strains in our collection was studied in detail (Figure 2.7). Each strain that had an ICE insertion in the *prfC3* gene in its genome was manually examined for the presence or absence of antibiotic resistance cassettes known to be variably present in the hot spots (variable regions) of this element (Table 2.3). Five different patterns were observed based on the antibiotic resistance genes possessed by the SXT ICE (Table 2.3). These patterns matched the clades in the maximum likelihood phylogenetic tree constructed on SNPs in the core regions of the SXT (Figure 2.8). Although the core SXT had a total length of ~60 kb, the number of SNPs called from this region were approximately three times the number of SNPs called from the total ~4 mb cholera genome. Thus, this SXT mutation rate is significantly different from the *V. cholerae* genomic backbone mutation rate, strongly indicating that the R391 family ICE or SXT must be evolving independently of the *V. cholerae* genomic pool. When the 5 SXT tree clades are coloured differently and the SXT type information is superimposed onto the seventh pandemic tree, it is clear that SXT would have entered the seventh pandemic lineage on at least five occasions (Figure 2.8). Furthermore, when compared on the same scale, the diversity in SXT and the diversity in the seventh pandemic El Tor lineage are significantly different (Figure 2.8). From genome assemblies, the point of first likely acquisition in the seventh pandemic was determined, which was found to be at the point of transition from wave-1 strains being the dominant clinical isolates to those of wave-2 and wave-3. Its entry in the seventh pandemic lineage (Figure 2.3) was also dated. It is also important to note that isolates collected in Vietnam between 1995-2004 were the only wave-2 isolates from this time period that lacked SXT. When the genomic locus in these clones that marks the point of insertion of SXT in all other *V. cholerae* isolates was checked for signatures of insertion or excision of SXT, no remnants of this conjugative element were found.

SXT Antibiotic Resistance --> Strain Name	*floR*	*Aph*	*strAB*	*sulII*	*dhfR*	*tetAR*	*MerRTPCA*	*czcD*
A346_2__Bangladesh_1994	+	-	+	+	+	-	-	+
A346_1__Bangladesh_1994	+	-	+	+	+	-	-	+
B33_Mozambique_2004	+	-	+	+	+	-	-	+
1627_Mozambique__2005	-	-	-	-	+	-	-	-
1346_Mozambique__2005	-	-	-	-	+	-	-	-
1362_Mozambique_2005	+	-	+	+	+	-	-	+

MJ1236_Bangladesh_M__1994	+	-	+	+	+	-	-	+
MJ1485_Bangladesh_M__1994	+	-	+	+	+	-	-	+
4623__India_2007	-	-	+	+	+	+	-	-
4536___India_2007	-	-	+	+	+	+	-	-
4552__India_2007	-	-	+	+	+	+	-	-
4600___India_2007	-	-	+	+	+	+	-	-
4585__India_2007	-	-	+	+	+	+	-	-
4551__India_2007	-	-	+	+	+	+	-	-
4593__India_2007	-	-	+	+	+	+	-	-
4488___India_2006	-	-	+	+	+	+	-	-
4605_India_2007	-	-	+	+	+	+	-	-
4122_Vietnam_2007	-	-	+	+	+	+	-	-
4672_Bangladesh_2000	+	-	+	+	-	-	-	-
A383__Bangladesh_2002	-	-	-	-	-	-	-	-
MO10_India_1992	+	-	+	+	+	-	-	-
A330_India_1993	+	-	+	+	+	-	-	-
CIRS101_Bangladesh_2002	+	-	+	+	+	-	-	-
6180_Nairobi_2007	+	-	+	+	+	-	-	-
4660_Bangladesh_1994	+	-	+	+	+	-	-	-
6196_Nairobi_2005	+	-	+	+	+	-	-	-
6191_Nairobi_2005	+	-	+	+	+	-	-	-
V109__India_1990	+	-	+	+	+	-	-	-
MBN17__India_2004	+	-	+	+	+	-	-	-
MG116226_Bangladesh_M__1991	+	-	+	+	+	-	-	-
4519___India_2005	+	-	+	+	+	-	-	-
6201_Nairobi_2007	+	-	+	+	+	-	-	-
4675_Bangladesh_2001	+	-	+	+	+	-	-	-
4663_Bangladesh_2001	+	-	+	+	+	-	-	-
6195_Nairobi_2005	+	-	+	+	+	-	-	-
4339__India_2004	+	-	+	+	+	-	-	-
4646___India_2007	+	-	+	+	+	-	-	-
4642_India_2006	+	-	+	+	+	-	-	-
MBRN14__India_2004	+	-	+	+	+	-	-	-
4656_India_2006	+	-	+	+	+	-	-	-
4322_India_2004	+	-	+	+	+	-	-	-
PRL64__India_1992	+	-	+	+	+	-	-	-
A487_1___Bangladesh_2007	+	-	+	+	+	-	-	-
6193_Nairobi_2005	+	-	+	+	+	-	-	-
7687_Machakos_2009	+	-	+	+	+	-	-	-
7685_Machakos_2009	+	-	+	+	+	-	-	-
7682_Machakos_2009	+	-	+	+	+	-	-	-
7684_Machakos_2009	+	-	+	+	+	-	-	-
7686_Machakos_2009	+	-	+	+	+	-	-	-
4538__India_2007	+	-	+	+	+	-	-	-
A488_1___Bangladesh_2006	+	-	+	+	+	-	-	-
4662_Bangladesh_2001	+	-	+	+	+	-	-	-
4784_Tanzania_2009	+	-	+	+	+	-	-	-
MG116025_Bangladesh_M__1991	+	-	+	+	+	-	-	-
6210_Nairobi_2007	+	-	+	+	+	-	-	-

6215_Kakuma_2005	+	-	+	+	+	-	-	-
6212_Kakuma_2007	+	-	+	+	+	-	-	-
6214_Kakuma_2007	+	-	+	+	+	-	-	-
A131_India_1989	+	-	+	+	+	-	-	-
A130_India_1989	+	-	+	+	+	-	-	-
A488_2___Bangladesh_2006	+	-	+	+	+	-	-	-
V5__India_1989	+	-	+	+	+	-	-	-
A482_Djibouti_2007	+	-	+	+	+	-	-	-
A481_Djibouti_2007	+	-	+	+	+	-	-	-
A483_Djibouti_2007	+	-	+	+	+	-	-	-
A487_2___Bangladesh_2007	+	-	+	+	+	-	-	-
4679_Bangladesh_1999	+	-	+	+	+	-	-	-
4661_Bangladesh_2001	+	-	+	+	+	-	-	-
6194_Nairobi_2007	+	-	+	+	+	-	-	-
IDHO1'726__India_2009	+	-	+	+	+	-	-	-
6197_Nairobi_2007	+	-	+	+	+	-	-	-
2010EL_1786_Haiti_2010	+	-	+	+	+	-	-	-
2010EL_1798_Haiti_2010	+	-	+	+	+	-	-	-
2010EL_1792_Haiti_2010	+	-	+	+	+	-	-	-

Chloramphenicol (floR); Kanamycin (Aph); Streptomycin (strAB);
Sulfonamide (sulII); Trimethoprim (dhfR); Tetracycline (TetAR);
Mercury (MerRTPCA); Cobalt/Zinc/Cadmium (czcD);

Wave 2	Wave 3

"+" = Resistant	**"-" = Sensitive**

Table 2.3: Chart showing the presence or absence of antibiotic resistance encoding gene cassettes carried variably within the SXT present in the seventh pandemic *V. cholerae* in our collection.

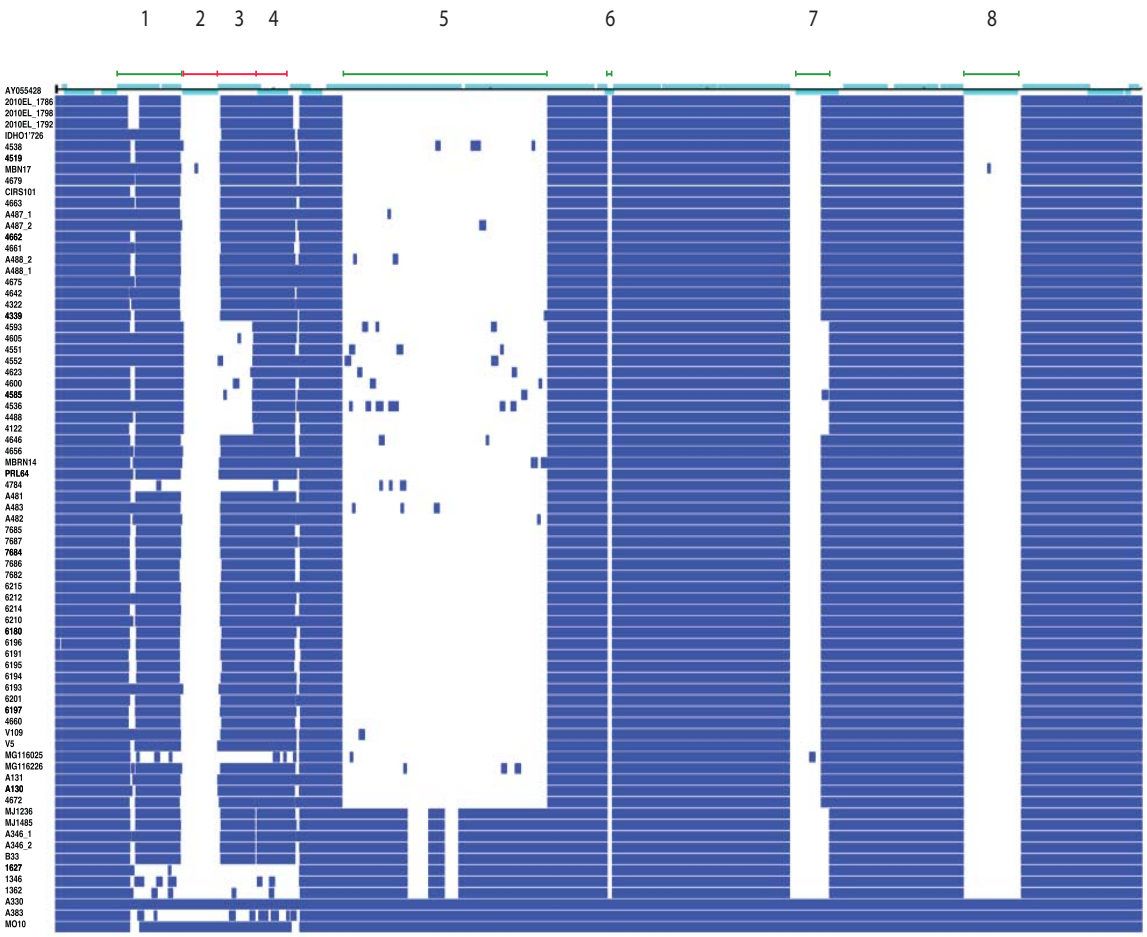


Figure 2.7: Plots showing presence (blue) or absence (white) information of the genes of the SXT of O139 strain MO10 (accession number AY055428) as reference in all the seventh pandemic strains positive for SXT/R391 ICE. The regions marked by green bars are variable and those marked with red bars are variable and encode antibiotic resistance (2, Trimethoprim; 3, Chloramphenicol; and 4, Streptomycin and Sulfonamide). Regions numerically marked are (as in Genbank): tnp - tnpB (1), dhfR18 - dcd (2), ‘tnpB - tnpB’ (3), strB - sulIII (4), s026 - s040 (5), s044 - s045 (6), s060 - s062 (7) and CDS - CDS (8).

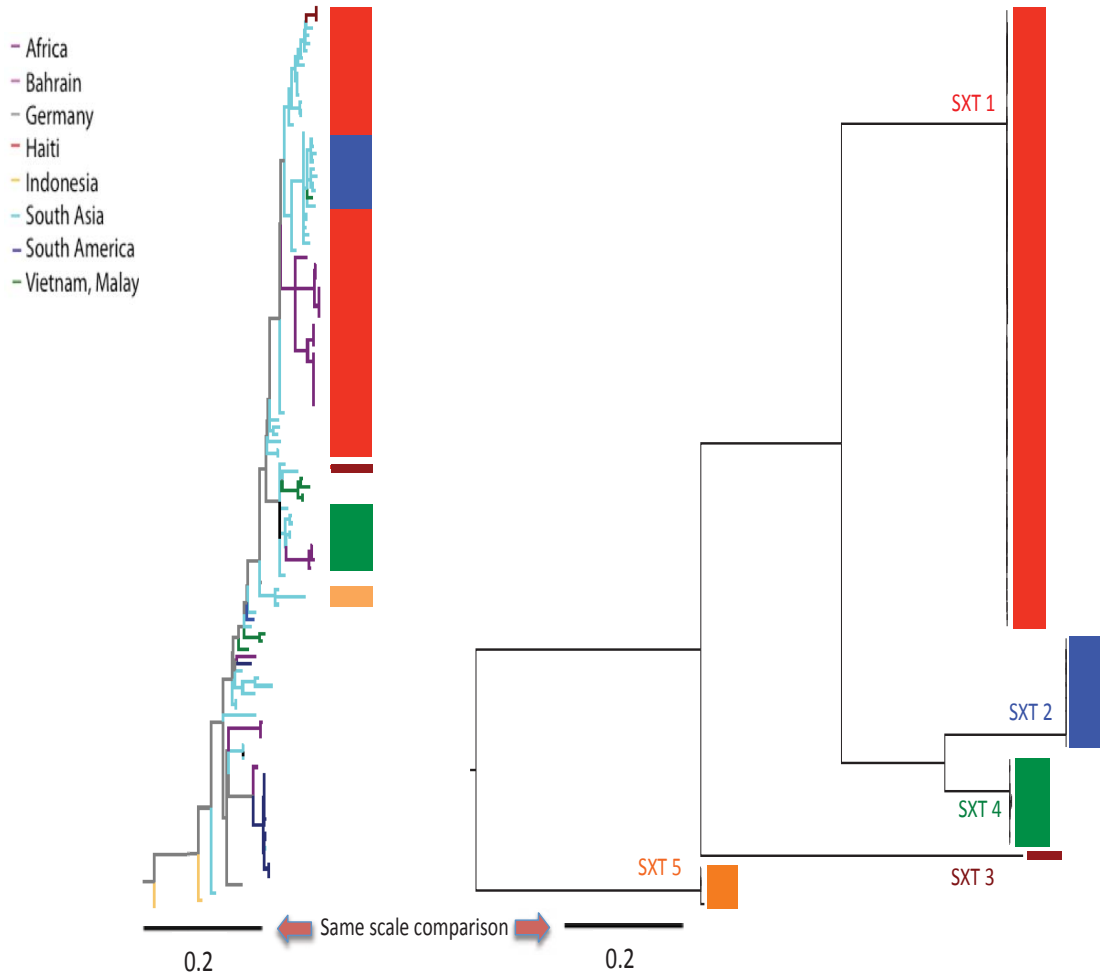


Figure 2.8: Comparison of the maximum likelihood trees of seventh pandemic lineage (left) and core SXT (right). The scales represent substitution per variable site and the colours of the blocks represent the SXT type 1 to 5. Core SXT of MO10, an O139 strain, was used to map the SXT positive isolates and O139 core SXT clade was used to root the SXT tree.

2.3.9 WASA-1 and other markers of the West African/South American (WASA) clade

The phylogenetic branch harboring the West African/South American (WASA) clade can be distinguished from all other *V. cholerae* by the acquisition of the novel VSP-2 (Davis and Waldor, 2003) gene island and a novel genomic island denoted here as "WASA-1" (described below, and in Table 2.3; note SNPs from these regions were not used to construct the seventh pandemic phylogeny). Strikingly, the Angolan isolate

A5 and all the South American isolates could be distinguished by just 10 SNPs. Based on the accumulation rate of 3.3 SNPs per year (Figure 2.4), the 3 year time period between the isolation of A5 and the oldest South American isolate A32 included in this study is consistent with previous studies that have suggested that cholera spread as a single epidemic after entering South America in 1991 (Lam, *et al.*, 2010).

Aside from the two key lateral gene transfer events (the acquisitions of CTX and SXT) described above, gene flux within the seventh pandemic lineage involved a further 155 genes (Figure 2.9). However, most of the flux was in the form of genomic islands restricted to the terminal nodes on the tree, except WASA-1 (Figure 2.10), GI's -14, -15, -v and -m which were found to be associated with particular lineages, suggesting that they are of limited relevance to the common biology of the group.

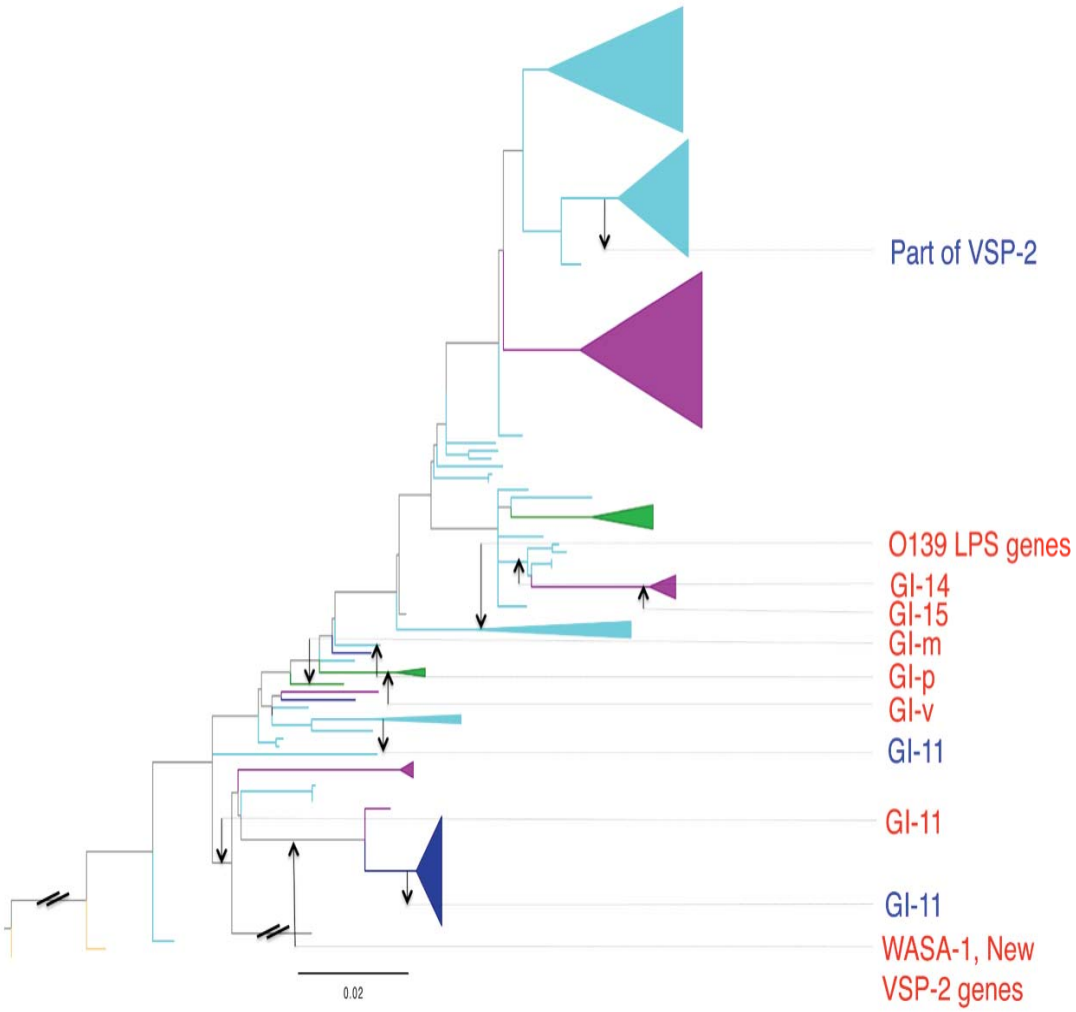


Figure 2.9: Seventh pandemic maximum likelihood tree with the gene flux plotted on the branches. Loci coloured red are insertions and those coloured blue are deletions.

The details of the genes carried on these genomic islands or region of differences are provided in the Table 2.4.

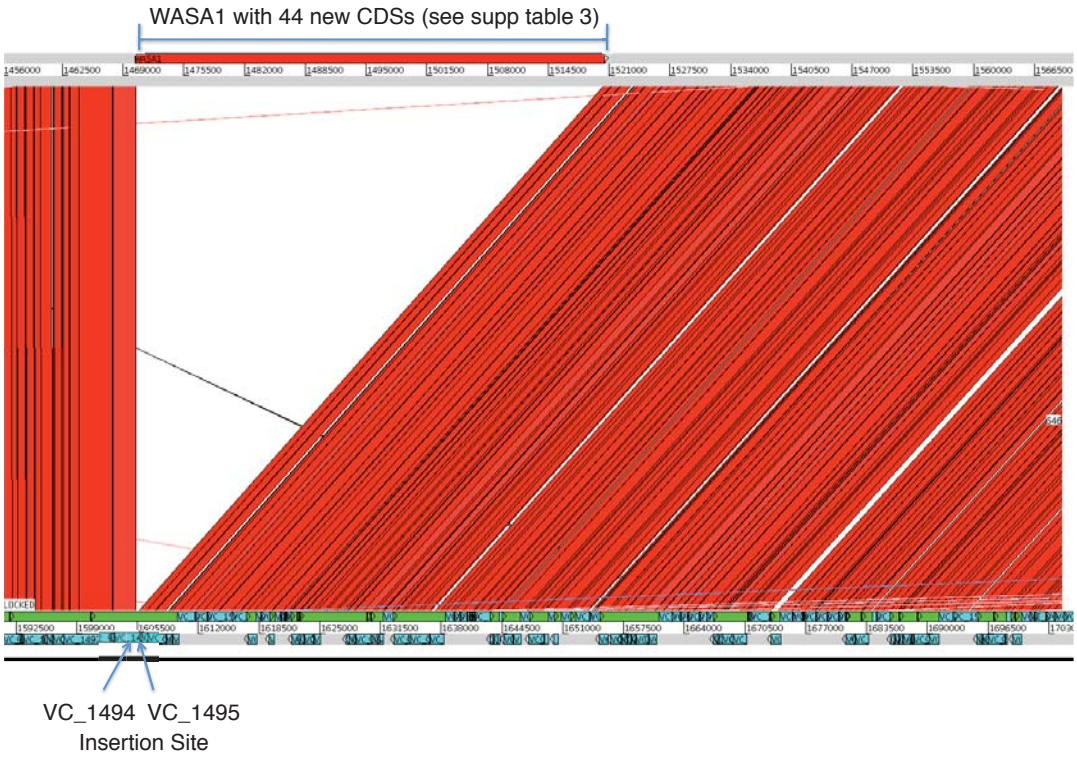


Figure 2.10: Insertion of WASA-1 within VC_1494 (product, aminopeptidase N). The genomes compared are A193 (top) and N16961 (bottom) and the insertion was present in all the WASA strains in our collection.

GI	Locus_tag (VC_)		1805	1806	1807	1808	1809	1810	1811	1812
GI-d	Function	PTS system	HP	HP	HP	HP	Putative transcriptional regulator	HP	Conserved HP	Conserved HP
			1813	1814	1815	1816	1817	1818	1819	1820
	Function		HP	Deoxyribodipyrimidine photolyase	Putative C-factor	HP	Sigma-54 dependent transcriptional regulator	HP	Aldehyde dehydrogenase	PTS system
			1821	1822	1823	1824	1825	1826	1827	1828
	Function		PTS system	PTS system	PTS system	PTS system	Transcriptional regulator	PTS system	Mannose-6-phosphate isomerase	Conserved HP
			1829	1830						
	Function		HP	HP						
	No of CDSs	26 (VC_1805-1830)								
GI-m	Locus_tag (New_)		1	2	3	4	5	6	7	
	Match to	Phage	Recombinase	Resolvase	NM	NM	NM	NM	NM	
	Next closest match		No	No	Transcriptional regulator	Exoribonuclease YhaM		No	No	
			8	9						
	Match to		Hypothetical Oxidoreductase	PUP						
	Next closest match		No	No						
	No of CDSs	9								
GI-v	Locus_tag (New_)		1	2	3	4	5	6		
	Match to	Phage	Integrase	PUP	Transposase	PUP	Putative DNA-binding protein	Putative DNA-binding protein		
	Next closest match		No	No	No	No	No	No	No	
	No of CDSs	6								
GI-p	Locus_tag (New_)		1	2	3	4				
	Match to	Phage	PUP	PUP	Protein SERAC1 of Erwinia Site specific recombinase					
	Next closest match		No	No	No	No	No			
	No of CDSs	4								
WASA-1	Locus_tag (New_)		1	2	3	4	5	6	7	8
	Match to	Phage	NM	NM	NM	PUP	PUP	PUP	NM	Phage tail tape measure protein
	Next closest match		No	No	No	No	No	No	No	No
			9	10	11	12	13	14	15	16
	Match to		PUP	Phage tail tape measure protein	PUP	PUP	NM	PUP	PUP	PUP
	Next closest match		No	No	No	No	No	No	No	No
			17	18	19	20	21	22	23	24
	Match to		PUP	Phage portal protein HK97	Phage Terminase	PUP	PUP	PUP	PUP	Prophage LP3 protein 12
	Next closest match		No	No	No	No	No	No	No	No
			25	26	27	28	29	30	31	32
	Match to		NM	NM	PUP	PUP	PUP	PUP	PUP	Exonuclease
	Next closest match		No	No	No	No	No	No	No	No
			33	34	35	36	37	38	39	40
	Match to		PUP	DNA Polymerase	DNA Primase	NM	PUP	PUP	NM	PUP
	Next closest match		No	No	No	No	No	No	No	No
			41	42	43	44				
	Match to		DNA directed RNA Polymerase	NM	NM	PUP				
	Next closest match		No	No	No	Recombinase				
	No of CDSs	44								
GI-11	See Chun *et al.* 2009									
GI-14	See Chun *et al.* 2009									
GI-15	See Chun *et al.* 2009									
VSP-2	See Chun *et al.* 2009									

PUP	Putative Uncharacterized Protein
NM	No Match
HP	HP
GI	Genomic Island

Table 2.4: List of all the genomic islands found in the seventh pandemic lineage and potential functions of the genes carried by them.

2.3.10 Recombination

Interestingly, apart from CTX, the seventh pandemic L2 isolates showed relatively little evidence of acquisition or recombination either within or from outside of the tree (as described below). Based on the SNP distribution, 1956 out of 2053 SNPs were congruent with the tree, leaving 97 homoplasic sites that could be due to selection or recombination within the tree. Just 296 SNPs were predicted to be due to recombination from outside the tree. The only two branches where the SNP distribution suggested significant recombination were those leading to the West African/South American cluster (Figure 2.11) and the O139 serogroup (Chun, *et al.*, 2009; Hochhut and Waldor, 1999).

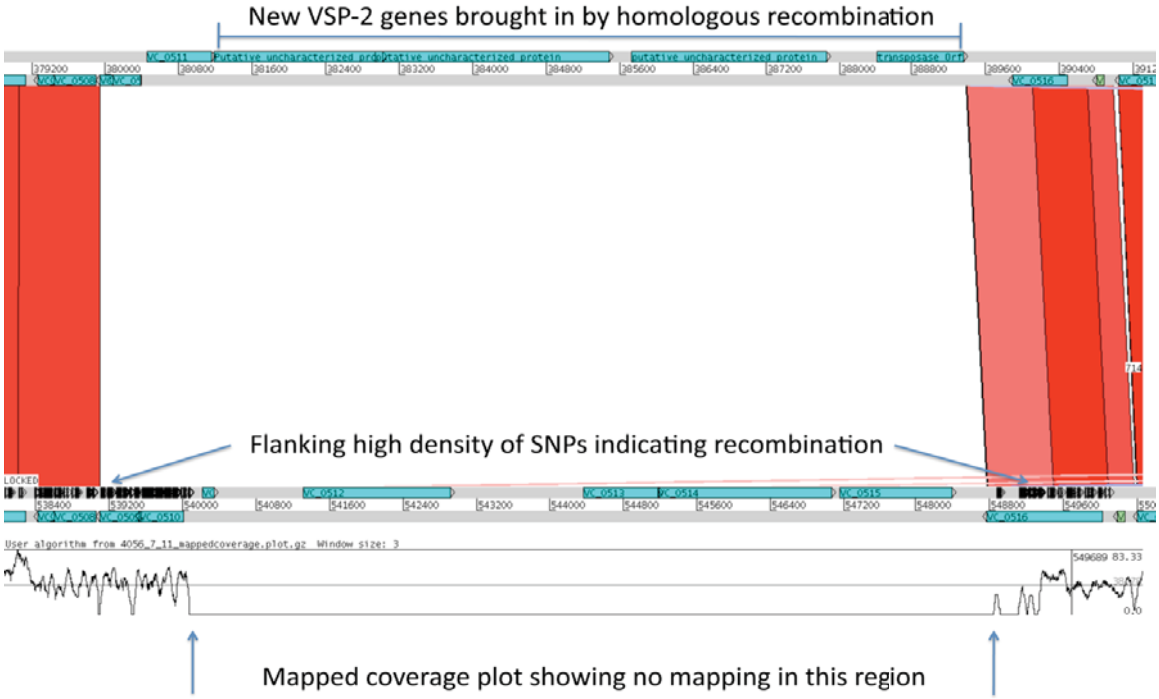


Figure 2.11: Recombination in the WASA cluster with homologous ends showing high SNP density and indicating likely recombination event from outside the tree. The genomes compared are A193 (top) and N16961 (bottom) where A193 represents the WASA cluster.

2.4 Conclusion and lessons from global phylogeny

The analysis of global *V. cholerae* population and the seventh cholera pandemic clearly suggests that classical and El Tor lineages are evolving independently and did not separate from a recent common ancestor. The seventh pandemic lineage is a clonal expansion from a single strain source and its spread has been in the form of at least three major overlapping but independent waves. One of the main contributing factors for the continuing success of current strains appears to be the acquisition of the multiple antibiotic resistance ICE element, SXT. Interestingly, the clinical use of antibiotics tetracycline and furazolidone for cholera treatment started in 1963 and 1968 respectively, ~15 years before the first acquisition of SXT according to our data.

It is clear from this data that the strains isolated from cholera affected parts of Haiti do cluster with the south Asian clade in wave-3. However, the number of SNP differences, even when using whole genome analysis, between the most closely related Indian and Bangladesh strains, is very low making any conclusions about the specific country of origin very difficult on sequence alone. In order to reach any such robust conclusions, sample collection from the bordering areas of neighboring nations and at parallel time points is required. The data also illustrates that intercontinental transmission in the form of pandemic waves or sporadic transmission due to travellers carrying *V. cholerae* are not ‘one off events’ in the history of the seventh pandemic as three independent but overlapping waves and four sporadic transmission events were identified in our limited collection. Such rapid long-range transmission events are consistent with human activity, as has also been shown in recent global transmission of clones of MRSA (Harris, *et al.*, 2010), and other bacteria.

3. Case studies on the regional evolution of *V. cholerae* O1 El Tor

NOTE: All the isolates were collected by our collaborators based in Pakistan, Kenya and South America. The DNA was sent to the Sanger Institute for sequencing by the sequencing pipeline teams and raw short read data was made available for the analyses. The work explained in this chapter details the regional phylogenetic analyses, which was done by me and therefore forms a part of my PhD thesis.

3.1 Introduction

Understanding the global spread of cholera provides a framework to study the transmission, spread and evolution of *V. cholerae* within a specified boundary. This is vital if we are to identify the foci to be targeted by the national, state and municipal level public health bodies. Well thought through actions at these foci could not only limit the spread within national, state, city or perhaps international boundary but such approaches could also prove to be effective in limiting the import, especially in non-endemic settings. Thus, the global and seventh pandemic *V. cholerae* phylogenetic framework (discussed in chapter 2) was used to analyse three important regional *V. cholerae* outbreaks. This chapter describes the phylogenetic and comparative genomic analyses of regional collections in the form of case studies.

The first study details countrywide epidemiological and microevolution investigation of a major cholera outbreak in Pakistan. In 2010, cholera was acknowledged, for the first time, as a serious public health threat in Pakistan despite numerous previous reports of sporadic outbreaks by different groups (Ahmed and Shakoori, 2002; Enzensberger, *et al.*, 2005; Jabeen, *et al.*, 2008). In late July and August 2010, record monsoon rainfall and the simultaneous glacier melt resulted in the worst flooding in the history of Pakistan, impacting an area of approximately 60,000 square miles and covering Khyber Pakhtunkhwa, Sindh, Punjab, Federal Administered Tribal Areas, Gilgit, Baltistan, Azad and Baluchistan provinces and displacing over 20 million people (see "Pakistan Floods: The Deluge of Disaster – Facts & Figures as of 15 September 2010"). The WHO reported 164 laboratory confirmed cases of cholera with the help of National Institute of Health (NIH) and allied departments in Pakistan

(WHO, 2011). However, the sources and routes of cholera infection and spread in Pakistan were not defined.

To understand the population dynamics and transmission of *V. cholerae* in flood affected and unaffected areas of Pakistan, the genomes of all clinical *V. cholerae* O1 El Tor reported during the flood disaster were sequenced and the data were compared to temporally representative genomes of a large global collection of *V. cholerae*. All the isolates were from clinical cases and collected over August to October in 2010, across North, East and South Pakistan.

Despite the lack of data on the impact of cholera in Pakistan prior to 2010, seasonal epidemics have occurred every year since then. Cholera is endemic in south-Asia (Sack, *et al.*, 2004) and the Bay of Bengal (Mutreja, *et al.*, 2011; Ramamurthy, *et al.*, 1993), where it is predominantly spread through contaminated food and water sources, often following civil unrest or natural disasters (Butler, 2010; Kondo, *et al.*, 2002). Pakistan particularly is at the risk of waterborne disease because it is largely an agricultural economy, with one of the most expansive water distribution systems in the world. These vast irrigation systems are largely dependent on the river Indus, which originates on the Northern slopes of the Kailash mountain range in India and runs North to South through the entire length of Pakistan with many tributaries including the Zaskar, the Shyok, the Nubra and the Hunza converging in the Northern region and flowing through the provinces of Ladakh, Baltistan and Gilgit. Therefore, understanding the routes of spread of cholera in Pakistan could provide the unprecedented opportunity to inform local and national public health agencies about specific foci where relevant action could limit the disease.

The second case study relates to a *V. cholerae* surveillance initiative in Kenya using clinical and environmental isolates collected over a period of 6 years (2005-2010) from the shores of Lake Victoria in the west to the Mombasa coastal region in the east and Nairobi and the surrounding urban or city areas in central Kenya. Sixty-six percent of the all cases of cholera reported world-wide between 1995 and 2005 were actually in sub-Saharan countries (Griffith, *et al.,* 2006) and since the report of the first official case of cholera on the continent in 1971 (Scrascia, *et al.*, 2006) at least 18 discrete outbreaks have been documented (Mohamed, *et al.,* 2012; Mugoya, *et al.*,

2008; Scrascia, *et al.*, 2009; Shapiro, *et al.*, 1999; Shikanga, *et al.*, 2009; Tauxe, *et al.*, 1995). From 2000 to 2006, the number of cases notified to the WHO each year ranged from 816 to 1,157. In 2007, a cumulative total of 625 cases resulting in 35 deaths were reported in four regions; Rift Valley (West Pokot, Turkana), Coast (Kwale), North Eastern (Garissa, Wajir, Mandera) and Nyanza (Kisumu, Bondo and Siaya). In addition from January–April 2008, in the Lake Victoria region of Kenya (Suba, Migori, Homabay, Rongo, Siaya, Kisumu, Bondo, Nyando, Kisii South), outbreaks resulted in 790 cases and 53 deaths. During the period January 2009–May 2010, cholera was reported in other regions including the coast with a total of 11,769 cases and 274 deaths (Mohamed, *et al.*, 2012; Shikanga, *et al.*, 2009). Recent work in Kenyan *V. cholerae* isolates identified 5 MLVA clonal complexes circulating in Kenya (Kendall, *et al.*, 2010; Mohamed, *et al.*, 2012). However, MLVA does not provide a phylogenetic context and therefore has limited utility in the tracking of outbreaks. Consequently, to understand how the *V. cholerae* causing the outbreaks in different Kenyan regions are related phylogenetically, this case study uses whole genome sequencing to establish how Kenyan *V. cholerae* are related to each other and the global *V. cholerae* population.

Finally, the third case study focused on Mexican isolates including historical *V. cholerae* isolates associated with the 1990s outbreaks in Latin American and recently collected isolates causing outbreaks between 2004-2010. It is likely that epidemic *V. cholerae* from Africa arrived in South America in the 1970s *via* Peru (see Chapter 2). The outbreak spread throughout Latin American, which had not experienced this disease for over a century (Franco, *et al.*, 1997). In June 1991, first cases of cholera were reported from a community on the banks of San Miguel river in Mexico and in the following five years ~43,000 cases were reported with incidence peaks in 1991, 1993 and 1995 (Borroto and Martinez-Piedra, 2000; Franco, *et al.*, 1997).

Based on traditional molecular genotyping techniques such as MLST and PFGE, multiple variants of the El Tor biotype have been identified across the cholera-affected regions of the world and it has recently been shown that classical, El Tor and a variant of El Tor biotype were all present in Mexico (Alam *et al.* 2010) between 1991-1997. However, little is known about their global and local phylogenetic relationship. Hence, to develop a high-resolution view of cholera in Mexico, the

genomes of 84 *V. cholerae*, including those from clinical cases, food and environmental samples were sequenced and analysed as part of this study. The collection spanned from 1991, the year the seventh pandemic of cholera first entered Mexico, to 2010 when cholera reached Haiti.

3.2 Bacterial isolates

3.2.1 Pakistan *V. cholerae* collection

38 *V. cholerae* were isolated from the clinical samples obtained from the patients that reported in hospitals in different provinces of Pakistan. The collection spanned the months of August through to October, 2010 i.e. from the start of the floods to the times when the floodwaters receded. All the isolates used in this study are listed in Table 3.1.

isolate	Isolation date	Town	Province	ENA Accession
F1DN4	October	Nowshera	KPK	ERR051745
F2D59	August	DIKhan	KPK	ERR051746
F4D48	August	DIKhan	KPK	ERR051748
F5D38	August	DIKhan	KPK	ERR051749
F7D30	August	DIKhan	KPK	ERR051751
F8D25	August	DIKhan	KPK	ERR051752
F11D4	August	DIKhan	KPK	ERR051755
F12D1	August	DIKhan	KPK	ERR051756
F14KPD3	October	Khairpur	Sindh	ERR051758
F15KTH7	October	Jamshoro	Sindh	ERR051759
F16KTH6	October	Jamshoro	Sindh	ERR051760
F17KTH4	October	Jamshoro	Sindh	ERR051761
F18KTH3	October	Jamshoro	Sindh	ERR051762
F19KTH2	October	Jamshoro	Sindh	ERR051763
S1KCH15	October	Karachi	Sindh	ERR051764
S2KCH17	October	Karachi	Sindh	ERR051765
S4KCH16	October	Karachi	Sindh	ERR051767
S5KCH10	October	Karachi	Sindh	ERR051768
S6KCH7	October	Karachi	Sindh	ERR051769
S7KCH20	October	Karachi	Sindh	ERR051770
S8KCH18	October	Karachi	Sindh	ERR051771

S9KCH9	October	Karachi	Sindh	ERR051772
S10P57	August	Peshawar	KPK	ERR051773
S12P76	August	Peshawar	KPK	ERR051752
S13P83	August	Peshawar	KPK	ERR051753
S14P9	August	Peshawar	KPK	ERR051777
S16HH1	October	Hyderabad	Sindh	ERR051779
S17HH3	October	Hyderabad	Sindh	ERR051780
S18HH4	October	Hyderabad	Sindh	ERR051781
S19HH5	October	Hyderabad	Sindh	ERR051782
S20HH14	October	Hyderabad	Sindh	ERR051783
S21HH15	October	Hyderabad	Sindh	ERR051784
S22HH17	October	Hyderabad	Sindh	ERR051785
S23HH18	October	Hyderabad	Sindh	ERR051786
S24RG6	August	Rawalpindi	Punjab	ERR051787
S25R22	September	Rawalpindi	Punjab	ERR051788
S26R24	September	Rawalpindi	Punjab	ERR051789
S27RG11	August	Rawalpindi	Punjab	ERR051790

Table 3.1: Table listing the *V. cholerae* used in the Pakistan cholera study. European Nucleotide Archives (ENA) accession numbers are provided and the sequence data can be downloaded using the open access EBI or NCBI databases.

3.2.2 Kenyan *V. cholerae* collection

The *V. cholerae* collection in the Kenyan case study (Table 3.2) span 2005-2010. The clinical isolates were obtained from patients that were diagnosed with cholera in hospitals and the environmental samples were collected from sources frequently visited by the local communities affected by the cholera outbreaks. The samples come from Lake Victorian region in Western Kenya, Nairobi region in the center and Mombasa and Kilifi region on the West coast of Kenya. A few samples were from the refugee camps in West Pokot and regions bordering Ethiopia.

Isolate (Environmental)	Isolation date (M/Y)	Location	ENA Accession
KNE7	4/2010	Kisumu	ERR117471
KNE3C	4/2010	Kisumu	ERR117473
KNE195	12/2009	Ahero	ERR117475
KNE59	4/2010	Kisumu	ERR117476

KNE081A	4/2010	HomaBay	ERR117477
KNE134	12/2009	SioPort	ERR117480
KNE102A	12/2009	Kisumu	ERR117481
KNE134B	12/2009	SioPort	ERR117482
KNE83	4/2010	Busia	ERR117483
KNE11B	4/2010	Kisumu	ERR117484
KNE170	12/2009	HomaBay	ERR117485
KNE083A	4/2010	HomaBay	ERR117488
KNE3G	4/2010	Kisumu	ERR117490
KNE17	2/2010	Kwale	ERR117491
KNE114	12/2009	Kisumu	ERR117494
KNE96	2/2010	Msambweni	ERR117496
KNE109A	12/2009	Kisumu	ERR117502
KNE53	4/2010	Kisumu	ERR117503
KNE85	4/2010	HomaBay	ERR117504
KNE109B	12/2009	Kisumu	ERR117505
KNE18	4/2010	Kisumu	ERR117506
KNE85C	2/2010	Msambweni	ERR117507
KNE096B	4/2010	Ahero	ERR117508
KNE10G	4/2010	Kisumu	ERR117518
KNE70	4/2010	HomaBay	ERR117520
KNE45	4/2010	Kisumu	ERR117528
KNE81	4/2010	HomaBay	ERR117532
KNE150	4/2010	HomaBay	ERR117535
KNE60	4/2010	Kwale	ERR117536
KNEXX	2/2010	HomaBay	ERR117544
KNEXC	12/2009	HomaBay	ERR117555
KNEXXH	12/2009	HomaBay	ERR117556
KNE056B_2	4/2010	Kisumu	ERR117561
KNE04C	4/2010	Kisumu	ERR117565
KNE104C	2/2010	Kwale	ERR117567
KNE98	4/2010	Ahero	ERR117568
KNE168	12/2009	HomaBay	ERR117569
VE1	4/2010	Kwale	ERR037738
VE2	2/2010	Mombasa	ERR037739
VE3	4/2010	Malindi	ERR03774
Isolate (Clinical)	**Isolation date (Y)**	**Location**	**ENA Accession**
6210	2007	Busia	ERR019290
6201	2007	Busia	ERR019291
6197	2007	Bahati	ERR019292
6196	2005	Bahati	ERR019293
6195	2005	Bahati	ERR019294
6194	2007	Bahati	ERR019295
6193	2005	Bahati	ERR019296

6215	2005	Bahati	ERR019297
6214	2007	Bahati	ERR019287
6191	2005	Bahati	ERR019288
6212	2007	Bahati	ERR019289
7682	2009	Kibera	ERR028066
7687	2009	Kibera	ERR028074
7686	2009	Kibera	ERR028075
7685	2009	Kibera	ERR028076
7684	2009	Kibera	ERR028068
KNC145	2010	Msambweni	ERR117571
KNC135	2009	Mathare	ERR117572
KNC151	2010	Kakuma	ERR117573
KNC8884	2010	Malindi	ERR117574
KNC56	2010	West pokot	ERR117577
KNC8880	2010	Malindi	ERR117578
KNC133	2007	Mathare	ERR117579
KNC64	2010	West Pokot	ERR117580
KNC8889	2010	Malindi	ERR117581
KNC149	2009	Makadara	ERR117582
KNC158	2010	West Pokot	ERR117583
KNC11	2010	West Pokot	ERR117586
KNC144	2010	West Pokot	ERR117587
KNC147	2010	Msambweni	ERR117588
KNC161	2010	West Pokot	ERR117590
KNC146	2010	Msambweni	ERR117591
KNC157	2009	Malindi	ERR117592
KNC1888	2007	Kwale	ERR117593
KNC156	2009	Malindi	ERR117594
KNC8885	2010	Malindi	ERR117595
KNC124	2009	Mathare	ERR117596
KNC155	2010	Kakuma	ERR117597
KNC143	2010	Kakuma	ERR117598
KNC8669	2010	Kisumu	ERR117599
KNC231	2009	Malindi	ERR117600
KNC205	2009	Thika	ERR117601
KNC1583	2009	Kibera	ERR117602
KNC241	2009	Thika	ERR117603
KNC206	2009	Thika	ERR117604
KNC233	2009	West Pokot	ERR117605
KNC8679	2008	Malindi	ERR117606
KNC1420	2009	Kakuma	ERR117607
KNC207	2009	Thika	ERR117609
KNC8675	2009	Daadab	ERR117610
KNC8572	2009	Daadab	ERR117612

KNC1509	2009	Kibera	ERR117613
KNC8673	2009	Kisumu	ERR114415
KNC8678	2009	Kisumu	ERR114416
KNC1709	2009	Kibera	ERR114417
KNC208	2009	Thika	ERR114418

Table 3.2: Table listing the environmental and clinical *V. cholerae* collection used in Kenyan surveillance case study. European Nucleotide Archives (ENA) accession numbers are provided and the sequence data could be downloaded using the open access EBI or NCBI databases.

3.2.3 Mexican *V. cholerae* collection

V. cholerae collected for this Mexican case study spanned 1991-2010. The isolates sequenced were from a historical collection associated with the 1990s Latin American cholera epidemic, from samples collected from patients between 1991-2010 and environmental sources such as river water, bottled water, food and sewage. Table 3.3 lists all the isolates that were part of this study.

Isolate	Isolation Date	Location	ENA Accession
7929_1991	1991	Chiapas	ERR163233
8012_1991	1991	Puebla	ERR163234
8022_1991	1991	Puebla	ERR163235
8204_1991	1991	Distrito Federal	ERR163236
8338_1991	1991	Tabasco	ERR163237
16974_1992	1992	Tabasco	ERR163238
19294_1992	1992	Nuevo León	ERR163239
33297_1993	1993	Guerrero	ERR163240
54267_1994	1994	Guanajuato	ERR163241
54328_1994	1994	Puebla	ERR163242
60452_1995	1995	Oaxaca	ERR163243
60483_1995	1995	Oaxaca	ERR163244
1401_2004	2004	Nayarit	ERR163245
1992_2004	2004	Nayarit	ERR163246
2006_2004	2004	Nayarit	ERR163247
2007_2004	2004	Nayarit	ERR163248
985_2007	2007	Sonora	ERR163249

2533_2007	2007	Hidalgo	ERR163250
3145_2007	2007	Nayarit	ERR163251
353_2008	2008	Nayarit	ERR163252
354_2008	2008	Nayarit	ERR163253
372_2008	2008	Michoacán	ERR163254
504_2008	2008	Nayarit	ERR163255
971_2008	2008	Nayarit	ERR163256
2908_2008	2008	Nayarit	ERR163257
3271_2009	2009	Tamaulipas	ERR163258
210_2010	2010	San Luis Potosí	ERR163259
211_2010	2010	San Luis Potosí	ERR163260
391_2010	2010	Tabasco	ERR163261
586_2010	2010	Tabasco	ERR163262
601_2010	2010	Nuevo León	ERR163263
667_2010	2010	Tabasco	ERR163264
819_2010	2010	Michoacán	ERR163265
2283_2010	2010	Puebla	ERR163266
2284_2010	2010	Puebla	ERR163267
2496_2010	2010	Sinaloa	ERR163268
2806_2010	2010	Distrito Federal	ERR163269
3056_2010	2010	Veracruz	ERR163270
204_2010	2010	San Luis Potosí	ERR163271
82	1998	Hidalgo	ERR108516
838	1999	Morelos	ERR108517
54	1999	Tabasco	ERR108518
1127	1999	Mexico	ERR108519
1876	1999	Mexico	ERR108520
909	1999	Mexico	ERR108521
85	2000	Chiapas	ERR108522
848	2000	Mexico	ERR108523
2710	2001	Mexico	ERR108524
2174	2001	Mexico	ERR108525
2370	2001	Mexico	ERR108526
2709	2001	Mexico	ERR108527
1354	2001	Mexico	ERR108528
644	2002	Mexico	ERR108529
1835	2003	Mexico	ERR108530
1582	2003	Mexico	ERR108531
1146	2004	Mexico	ERR108532

1148	2004	Mexico	ERR108533
1596	2004	Mexico	ERR108534
2006	2004	Nayarit	ERR108535
5032	2005	Nayarit	ERR108536
688	2006	Nayarit	ERR108537
1474	2007	Mexico	ERR108538
353	2008	Nayarit	ERR108539
EM- 0892	2002	Mexico	ERR108540
87211	1991	Mexico	ERR044778
116072	1991	Mexico	ERR044779
87258	1991	Mexico	ERR044780
116073	1991	Mexico	ERR044781
87151	1992	Mexico	ERR044782
116075	1992	Mexico	ERR044783
87397	1993	Mexico	ERR044784
87667	1993	Mexico	ERR044785
87662	1993	Mexico	ERR044786
87406	1994	Mexico	ERR044787
87409	1995	Mexico	ERR044788
97639_1	1995	Mexico	ERR044789
93154	1996	Mexico	ERR044790
95430	1997	Mexico	ERR044791
95409	1997	Mexico	ERR044792
95412	1997	Mexico	ERR044793
Mex1	1991	Mexico	ERR042753
Mex6	1992	Mexico	ERR042754
Mex15	1997	Mexico	ERR042755
Mex16	1997	Mexico	ERR042756

Table 3.3: Table listing the environmental and clinical *V. cholerae* collection used in the Mexican case study. European Nucleotide Archives (ENA) accession numbers are provided and the sequence data could be downloaded using the open access EBI or NCBI databases.

3.3 Results and discussion

3.3.1 Whole genome phylogeny of 2010 Pakistan flood *V. cholerae*

Whole genome sequences of the 38 *V. cholerae* O1 El Tor from Pakistan were determined using the Illumina sequencing platform. A high resolution phylogenetic tree based on SNPs was constructed by mapping the sequence reads to the completed reference genome sequence of *V. cholerae* O1 El Tor strain N16961 (isolated in Bangladesh in 1975, accession No. AE003852-3). SNPs were only counted from the non-mobile and non-repetitive parts of the genome. To determine if these isolates fitted into the global *V. cholerae* El Tor phylogeny, the genomes of 146 previously published *V. cholerae* O1 El Tor from different parts of the world (see Chapter 2) were included in the tree. The consensus tree showed that all the isolates from Pakistan fell within two contemporary sub-clades (Pakistan sub-clade 1 or PSC-1 and Pakistan sub-clade 2 or PSC-2), both of which branched from different positions within the third transmission wave of the seventh pandemic lineage (Mutreja, *et al.*, 2011) (Figure 3.1). After removing genomic recombination sites, the variation in the El Tor global phylogeny could be defined by 1,826 variable genomic sites. PSC-1 and PSC-2 harboured only 12 and 22 SNPs respectively that distinguished them from their third wave ancestors. Thus, within each sub-clade the isolates were very closely related, with only 4 SNPs within PSC-1 and 76 SNPs amongst the PSC-2 isolates.

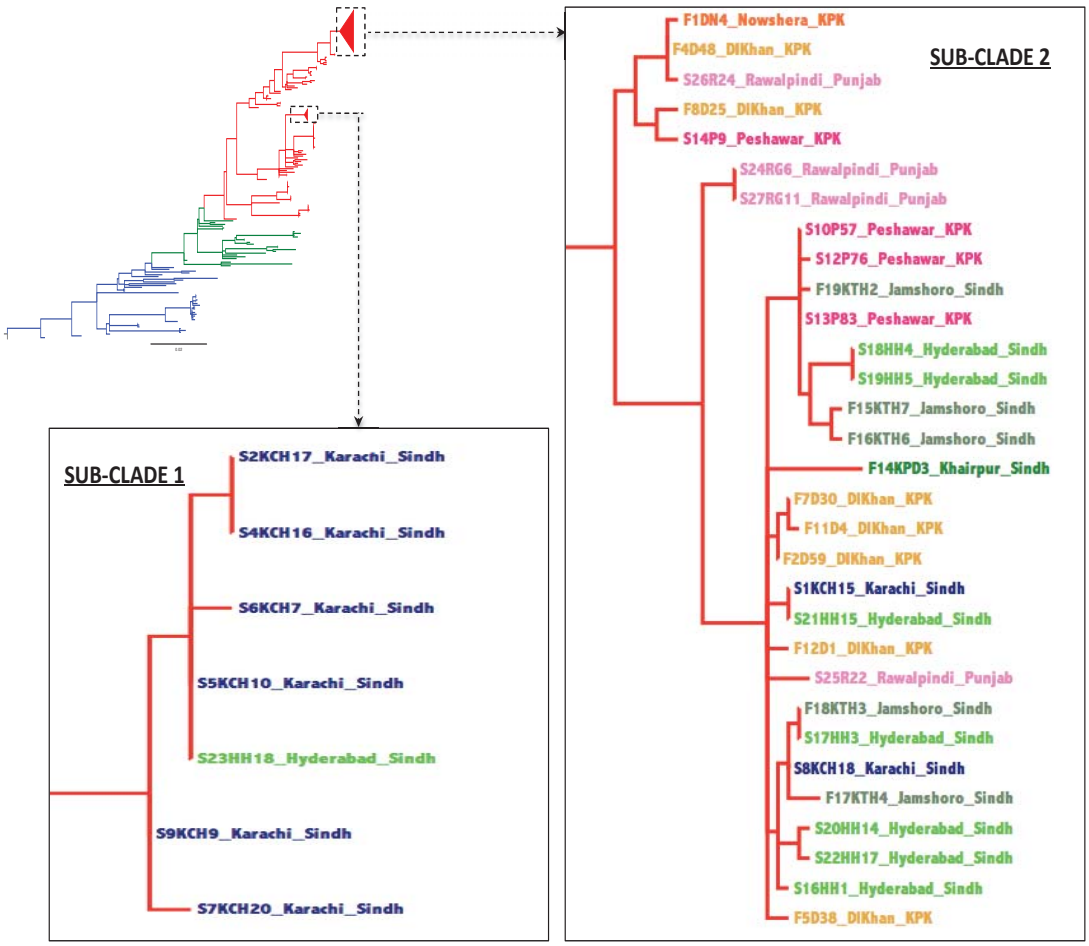


Figure 3.1: Maximum Likelihood phylogenetic tree based on the SNP variation in the Pakistan isolates, showing the relative position of the Pakistani *V. cholerae* O1 El Tor in the wave-3 of the seventh pandemic lineage. The blue, green and red colours of the branches in the tree represent wave-1, 2 and 3 respectively. Different colours of the sub-clade 1 and 2 isolates represent different locations where they were isolated from.

When the data for all the Pakistani isolates were plotted on to the map we noticed that all the PSC-1 isolates were derived from cholera cases located in the coastal city of Karachi (6/7) and Hyderabad (1/8), whereas the PSC-2 isolates were sourced from a wider geographical region comprising flood affected and non-flood inland regions (30/31) with one case from Karachi (Figure 3.1, 2A). As genomic variation in the seventh pandemic El Tor *V. cholerae* occurs at a clock-like rate (Mutreja, *et al.*, 2011), root-to-tip distances of the PSC-1 and PSC-2 isolates were determined and the order was plotted onto a graph. The isolation date and geographical location information was superimposed (Figure 3.2B).

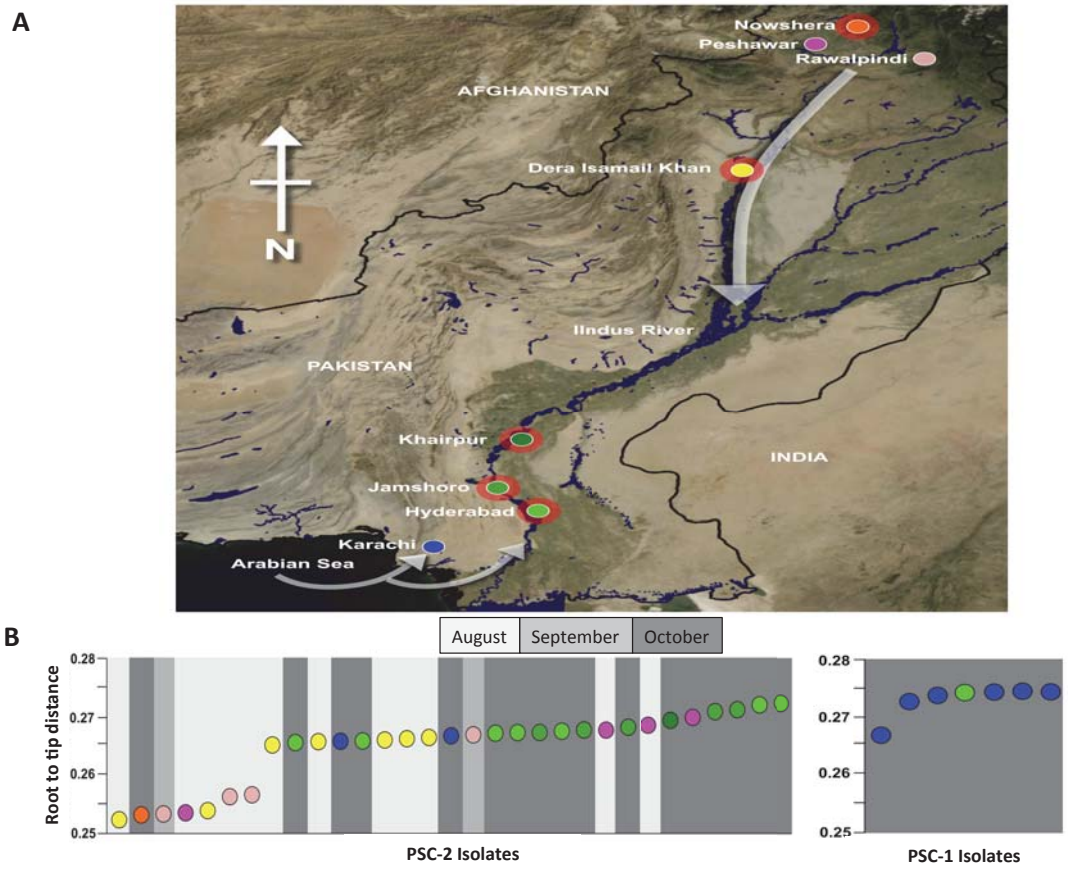


Figure 3.2: A) North orientated map of Pakistan indicating the eight locations where the *V. cholerae* O1 El Tor from this study were isolated (shown by individual coloured circles, red outer shading highlights the five locations that experienced flooding). White arrows show the hypothesised directions of the spread of cholera in Pakistan; B) Cumulative root-to-tip distances of PSC-2 and PSC-1 *V. cholerae* O1 El Tor isolates arranged in ascending order. Each coloured circle corresponds to an individual *V. cholerae* O1 El Tor isolate and colours relate with the locations shown in Figure 3.2A.

3.3.2 Evidence for a strict *V. cholerae* molecular clock in Pakistan

The data shown in Figure 3.2B strongly suggested that the PSC-2 isolates sourced from the northern regions of Pakistan show a shorter root-to-tip distance and were isolated earlier in time compared to the isolates collected from the southern regions of Pakistan, which showed longer branch lengths and were relatively distant from the root of the PSC-2 clade (Figure 3.2B). To statistically confirm the pattern seen in

Figure 3.2B, the mutation rate was calculated by plotting a linear regression curve (Figure 3.3) between the isolation date of each sample and the root to tip distance (R^2 = 0.27, $p < 0.001$). The rate of change of root to tip distance information was used to calculate the SNP acquisition rate in the Pakistan isolates and it was found to be a rate of 0.288 SNPs/month (3.4 SNPs/year). This is in accordance with the previous estimations of 3.3 SNPs/year inferred for the global seventh pandemic lineage (Chapter 2). This suggests that the isolates of PSC-2 clade, after entering from North Pakistan, travelled southwards with the Indus river and seeded cholera outbreaks along the course of the river, wherever the flood waters managed to break the banks (indicated by north to south arrow in Figure 3.2A). Interestingly, the Pakistani *V. cholerae* that clustered as a separate clade in PSC-1 were restricted to the southern or coastal parts of Pakistan and were all isolated later in time (Figure 3.2B). This suggests that there were two independent cholera epidemics with different geographical origins, occurring at the same time in Pakistan. It may be that PSC-1 was introduced into Pakistan separately, most likely from Arabian Sea (shown by south to north arrows in Figure 3.2A), and this clade had a limited range of spread during the flooding period because of the direction of the water current of Indus River draining into the sea. More *V. cholerae* collected at a later time point in the floods would be needed to confirm if PSC-1 really spread through to Central and Northern Pakistan with travellers or *via* transport of sea food from coastal Pakistan to its mainland regions. A future study to follow my PhD is already planned to address this question.

Furthermore, the earlier isolates of PSC-2, displaying shorter root-to-tip distances, were isolated in closer proximity to the source of the Indus River and were mainly isolated from Peshawar, Nowshera, Rawalpindi and DI Khan in the north of Pakistan. Conversely, isolates in October were from the southern regions of Pakistan, namely Khairpur, Jamshoro, Hyderabad and Karachi (Figure 3.2B). Another root-to-tip distance plot of the PSC-2 sub-clade against the distance from the source of the river Indus confirmed this association (R^2=0.35, $p < 0.001$; Figure 3.4). The observed pattern was consistent with the origins and progression of the floods, which began in Peshawar in late July and followed the course of the Indus river southwest passing Nowshera, DI Khan, Khairpur, Jamshoro and Hyderabad in August.

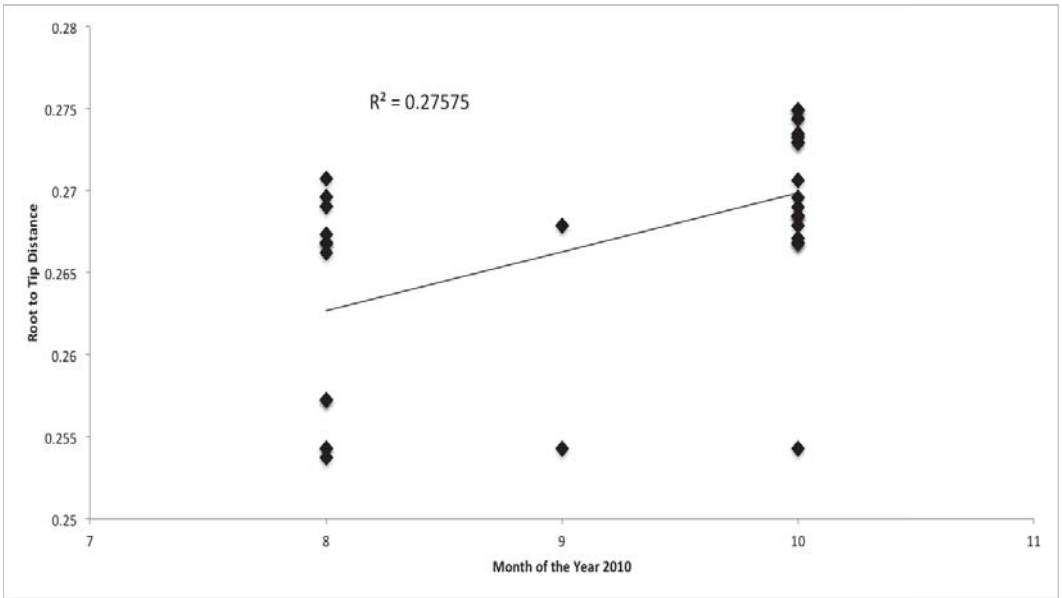


Figure 3.3: A scatter plot of root-to-tip distance *vs.* date of isolation for PSC-1 and PSC-2 combined. The R^2 value represents the correlation between root to tip distance of the Pakistan isolates and their time of isolation.

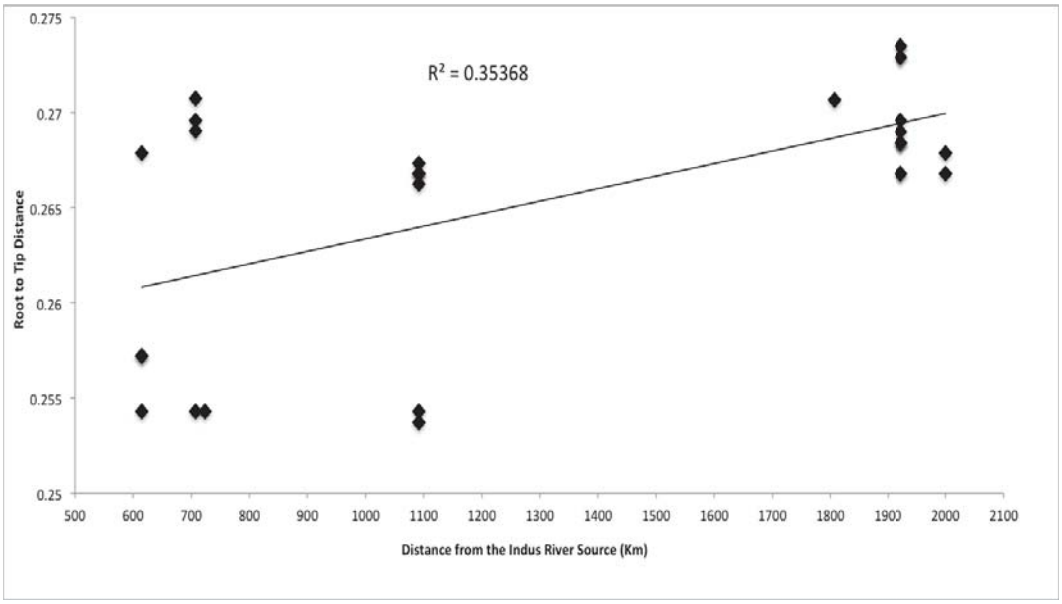


Figure 3.4: A scatter plot of root-to-tip distance *vs.* distance from river source for PSC-2. The R^2 value represents the correlation between root to tip distance of the Pakistan isolates and their distance from the source of the Indus river into Pakistan.

3.3.3 Sub-clade signature deletions within the genomes of Pakistan *V.*

cholerae

Further sequencing analysis showed that Pakistan sub-clades could be distinguished from other El Tor *V. cholerae* by sub-clade specific deletions identified in their genomes, particularly in the two pathogenicity islands: Vibrio pathogenicity island-1 and Vibrio seventh pandemic island-2 (VPI-1 and VSP-2). These genomic islands are known to encode functions that impact on the relative transmissibility and virulence of epidemic *V. cholerae*. All the PSC-1 isolates had a unique three-gene deletion in the VPI-1 pathogenicity island (VC_0819-0821), which included *aldA* (aldehyde dehydrogenase), *tagA* (a mucinase) and a predicted coding sequence encoding a hypothetical protein. TagA plays a role in host cell surface modification, an important step in preparation of host cell for bacterial attachment (Szabady, *et al.*, 2011) and this deletion may affect the virulence and transmissibility of the PSC-1 isolates. To my best knowledge, this deletion has not been previously reported, however, a deletion of the entire VPI-1 region was reported in an isolate from a patient in the US who had a history of travelling to Pakistan (Reimer, *et al.*, 2011). Additionally, a four-gene deletion within the VSP-2 (VC_0495-0498) was noted in PSC-1, which has previously been identified in El Tor *V. cholerae* responsible for cholera outbreaks in Bangladesh in 2008 (Chin, *et al.*, 2011). PSC-2 isolates, in contrast, had an 18-gene deletion (VC_0495-0512) in VSP-2, comparable to some of the most recently characterized strains of wave-3 of El Tor *V. cholerae* O1, including those from the Haitian cholera outbreak in 2010 (Chin, *et al.*, 2011) and from South East China in 2005 (Chin, *et al.*, 2011; Pang, *et al.*, 2007). VPI-1 is intact in PSC-2 isolates, except for a frame-shift mutation in the accessory colonization factor gene, *acfC* (VC_0841). These VSP-2 deletions are consistent with the position of the Pakistan sub-clades on the seventh pandemic phylogenetic framework discussed in Chapter 2. However, the relative impact of these deletions on *V. cholerae* pathogenesis and relative transmissibility remains to be evaluated.

3.3.4 Diversity within *V. cholerae* circulating in Kenya

In a surveillance study spanning years 2005-2010, 57 *V. cholerae* isolates were obtained from clinical cases of cholera, and 40 isolates were collected from environmental sources. The environmental isolates were derived from nine study sites

in Kenya where the pH of water was ranging from 4 to 9.7. The demography of the sample sites ranged from waters with algal blooms to areas where regular household and farming activities were performed. Water samples, plant materials and sediments from unprotected boreholes, wells, rivers, and surface runoffs were collected in the following towns bordering the Indian Ocean coast (Mombasa, Malindi, Kilifi and Kwale). Four sites were sampled along the shore of Lake Victoria (Kisumu, Siaya, Homa Bay and Kendu Bay) and three district towns were sampled in Western Kenya (Busia, Vihiga and Kakamega). A map showing the locations from where the Kenyan isolates used in this study were sourced is shown in Figure 3.5.

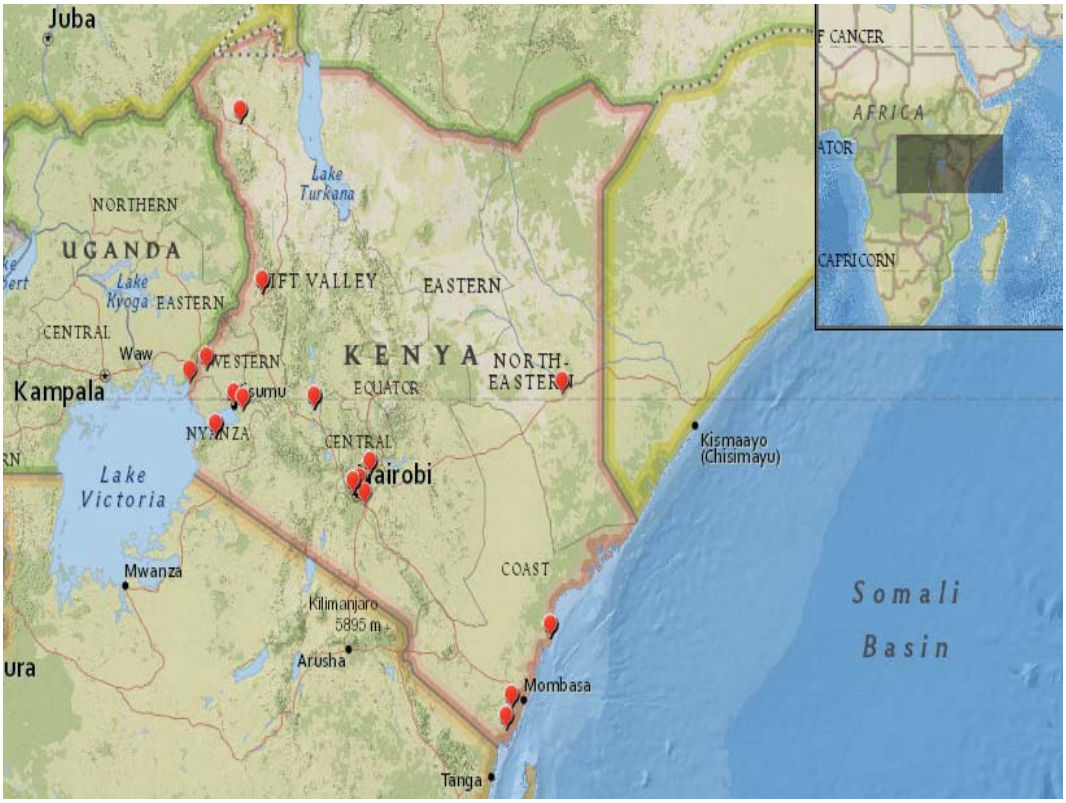


Figure 3.5: The map of Kenya showing the sites from where the isolates of *V. cholerae* were obtained. The inset shows the Kenyan region in context of Africa and each red balloon indicates a sampling site.

All the Kenyan O1 *V. cholerae* isolates, whether clinical or environmentally sourced, were biotyped and serotyped as Inaba, whereas some of the environmental isolates were found to be non-O1. When tested using a spectrum of antibiotics, all the clinical

isolates were resistant to multiple antibiotics, including nalidixic acid, trimethoprim, sulphamethoxazole, streptomycin and furazolidone. In contrast, 66% of the environmental isolates were resistant to sulphamethoxazole, 15% to furazolidone, 56% to ampicillin and 5% to trimethoprim. Interestingly, all the clinical isolates were fully susceptible to ampicillin. In addition both clinical and environmental *V. cholerae* isolates were also fully susceptible to tetracycline, cefuroxime, chloramphenicol and ciprofloxacin. This contrasts to some extent with previous Kenyan studies (Mwansa, *et al.*, 2007), in which 8% and 3% of clinical and environmental isolates, respectively, taken from around Lake Victoria were resistant to tetracycline. This resistance trend has some similarities to the picture emerging in endemic areas of Bangladesh where isolates are now uniformly resistant to trimethoprim/sulfamethoxazole and furazolidone with temporal variation in tetracycline and erythromycin resistance (Rashed, *et al.*, 2012). *V. cholerae* O1 resistant to tetracycline have previously been reported in Zambia (Mwansa, *et al.*, 2007) in the 1990s, but those isolated from Ethiopia (Scrascia, *et al.*, 2009) and Somalia (Scrascia, *et al.*, 2009) in the same period were found to be susceptible to this antibiotic.

3.3.5 The phylogeny of Kenyan *V. cholerae* based on whole genome sequences

DNA prepared from all the Kenyan clinical and environmentally derived isolates isolates were sequenced and the data generated was compared to previously published *V. cholerae* sequences (Hendriksen, *et al.*, 2011; Mutreja, *et al.*, 2011). The initial consensus phylogenetic tree generated from this data (Figure 3.6A) showed that 27 of the environmental isolates clustered well outside of the seventh pandemic lineage and differed by more than 50,000 SNPs from the reference N16961 El Tor O1 *V. cholerae*. Further, only 49% to 89% of the sequence reads of these 27 isolates mapped onto the *V. cholerae* El Tor reference genome making them markedly distinct from *V. cholerae* O1 El Tor lineages. However, the remaining 13 environmentally sourced isolates, which were also phenotypically serotyped O1 Inaba, clustered with other O1 El Tor seventh pandemic isolates. 98% of the sequence reads from these isolates mapped onto the N16961 El Tor reference genome, differing by only ~250 SNPs from this reference. The genomes of these isolates also harboured key signature genomic loci including VSP-1 and 2 and the SXT multiple antibiotic resistance

associated loci, characteristic of wave-3 seventh pandemic isolates (Mutreja, *et al.*, 2011). This data unequivocally confirms that these environmentally derived *V. cholerae* O1 isolates are members of the wave-3 of the seventh pandemic *V. cholerae* El Tor phylogeny (Mutreja, *et al.*, 2011).

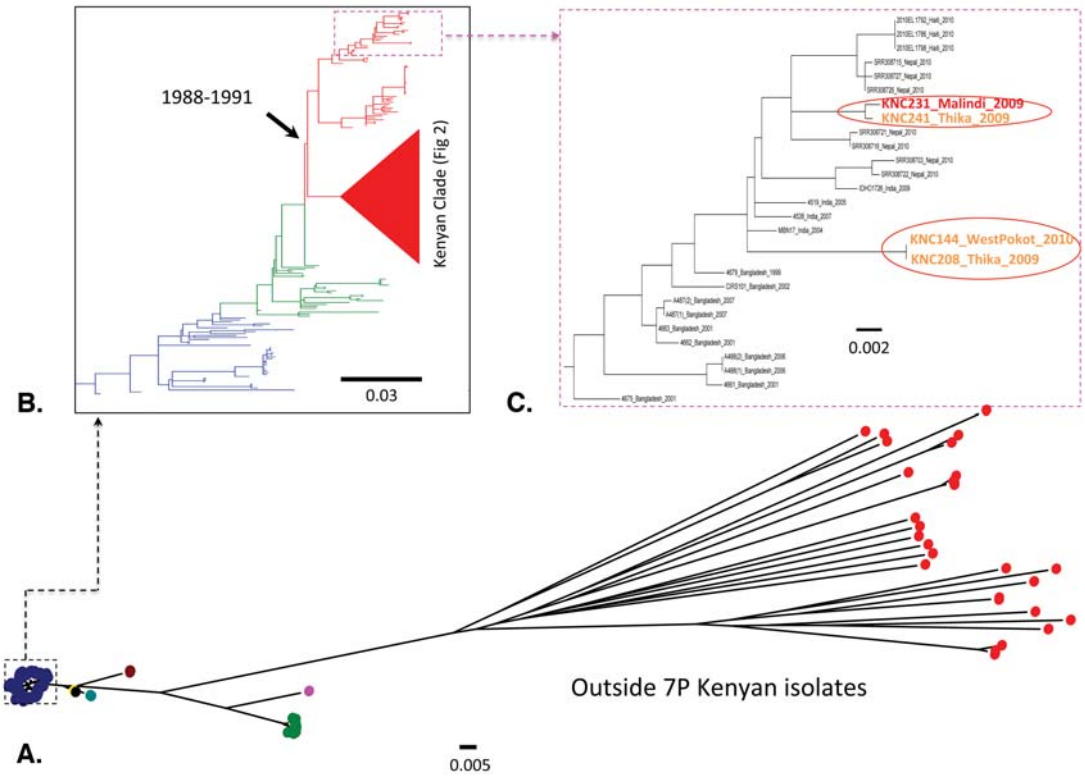


Figure 3.6: A) A Maximum Likelihood phylogenetic tree of *V. cholerae* based on SNP variation. The 6 major O1 clinical groups are shown in this tree with the seventh pandemic El Tor (in blue circles) and classical lineage (in green circles). Environmental non-O1/O139 from Kenya are represented as red circles. B) A maximum-likelihood phylogenetic tree of the seventh pandemic strains after exclusion of likely recombination events. The date range shown on the node is the BEAST estimated time when the seventh pandemic wave-3 cholera entered Kenya. M66, a previously published pre-seventh pandemic isolate, was used as an outgroup to root the tree. Blue, green and red coloured branches represent wave 1, 2, 3 and the red shaded clade represents dominant Kenyan respectively. C) A maximum likelihood phylogenetic sub-tree shows Kenyan isolates clustering with the south-Asian isolates. All the scales are given as the number of substitutions per variable site.

For a more detailed understanding of the phylogeography of the El Tor isolates from Kenya, a genome-wide SNP based phylogenetic tree of the El Tor seventh pandemic lineage was constructed (Figure 3.7). High density SNPs and any variation that could be deemed a consequence of recombination were removed using the method of Croucher *et al.* (Croucher, *et al.*, 2011). This tree was based on 1828 variable sites. 53 O1 serogroup Kenyan isolates clustered within the wave-3 of the global seventh pandemic lineage (Figure 3.6B), where 49 isolates formed an exclusive Kenyan clade alongside 17 previously published Kenyan isolates (Figure 3.6B, 7A). Interestingly, four *V. cholerae* O1 isolates (KNC231, KNC241, KNC144 and KNC208) clustered in distinct positions within a clade of isolates from south-Asia (Figure 3.6C), raising the possibility that these isolates could have been brought into Kenya independently by travellers from south-Asian sub-continent.

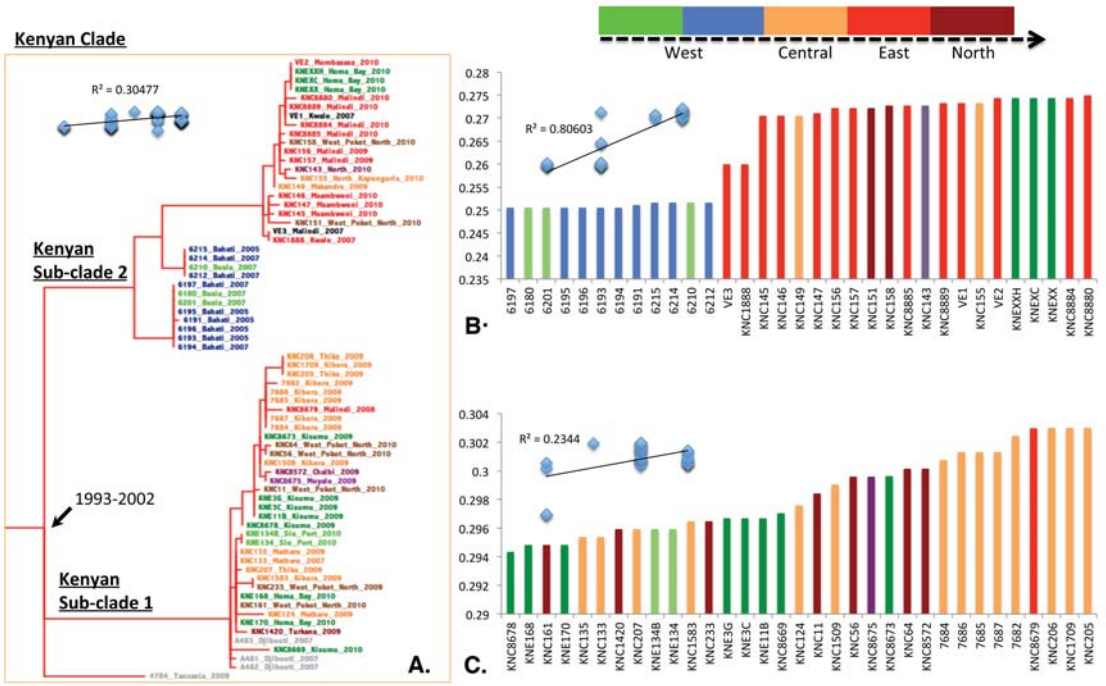


Figure 3.7: A) the exclusive Kenyan clade with two sub-clades, KSC-1 and KSC-2, from within the seventh pandemic maximum likelihood phylogenetic tree. The date on the node represents tMRCA of KSC-1 and 2. B) and C) root to tip distances of isolates of sub-clades KSC-2 and KSC-1 arranged in increasing order of magnitude. The R^2 values and the linear regression curves are based on root-to-tip distance *vs.* time (years) on vertical and horizontal axes respectively. The colours of the isolates in

7A and bars in 7B and 7C indicate the locations where the sample was collected. The root to tip distance for strains from Djibouti and Tanzania in KSC-1 are not provided in 7C.

To gain further insight into the temporal and spatial distribution of the Kenyan lineages the phylogeny of the Kenyan isolates was determined using a Bayesian analytical tool for mapping isolates against time. This data showed that wave-3 isolates of the seventh cholera pandemic entered Kenya around 1988-1991, a refinement on the previous estimates (Chapter 2), which were based on fewer Kenyan isolates (1989-1997) (Mutreja, *et al.*, 2011). A linear regression analysis was performed on the Kenyan clade by plotting the root-to-tip distance of each isolate against time of isolation. This data, consistent with the previous findings for the seventh pandemic lineage, showed that Kenyan cholera isolates evolved in a clock-like manner (Figure 3.7A, 7B, 7C). This refined analysis further subdivided the dominant Kenyan clade into two sub-clades, designated Kenyan sub-clade 1 (KSC-1) and Kenyan sub-clade 2 (KSC-2) (Figure 3.7A). Most of the isolates collected between 2005 and 2010 cluster within one of these two sub-clades. The most recent common ancestor for these two sub-clades was estimated to have emerged between 1993-2002 (Figure 3.7A).

Similar to the Pakistan isolates (Figure 3.1), there was evidence of regional clustering for clinical isolates within sub-clades KSC-1 and KSC-2 (Figure 3.7B, 7C). For example, with few exceptions, isolates from the Nairobi region fell within the KSC-1 sub-clade (Figure 3.7B) while most isolates (clinical and O1-positive environmental) from the Indian Ocean coast region (Mombasa, Msambweni, Kwale and Malindi) fell within KSC-2, as did those from Busia on the Kenya-Uganda border. Other isolates from the Lake Victoria region of Kisumu and Sio-Port clustered in KSC-1 whilst isolates from the Homa-Bay area were distributed in both KSC-1 and KSC-2. The isolates from the semi-arid region of West Pokot in Northern Kenya were also distributed in both the sub-clades, as were those from environmental sources near Lake Victoria (Figure 3.7B, 7C).

There was a strong correlation between root-to-tip distance and time for KSC-2 (R^2 = 0.8). The phylo-geographic analysis of KSC-2 is consistent with the notion that

cholera may emerge from in and around Lake Victoria and spread to the central and eastern parts of Kenya (Figure 3.7B). However, the same correlation for KSC-1 was weak ($R^2 = 0.2$) and was phylo-geographically inconclusive.

Currently, it is not known how cholera entered Kenya during this period but since the two sub-clades KSC-1 and KSC-2 were clearly distinct and we were able to identify travel linked outliers on the *V. cholerae* El Tor tree, this suggests that there were multiple introductions of the seventh pandemic *V. cholerae* into this country. Also, a hypothesis of how cholera is spreading and persisting within Kenya could also be drawn.

3.3.6 Genomic features of Kenyan O1 El Tor sub-clades

Short read data was used to perform *de novo* assembly of each isolate and a manual comparison of each assembled genome was made against the N16961 reference genome. All of the Kenyan O1 El Tor isolates possessed the Vibrio seventh pandemic islands 1 and 2 and possessed the site-specific insertion of the R391 family ICE/SXT multiple antibiotic resistance cassette. Uniquely, every isolate in the Kenyan clades KSC-1 and KSC-2 harboured a 4 gene (VC0495-VC0498) deletion in the VSP-2 island. Also, the travel linked Kenyan wave-3 isolates that did not cluster within the exclusive Kenyan clade but clustered with the south-Asian strains possessed an 18-gene (VC0495-VC0516) deletion characteristic of that south-Asian wave-3 clade (Mutreja, *et al.*, 2011).

All the Kenyan isolates that fell within wave-3 of the *V. cholerae* El Tor lineage harboured the R391-ICE/SXT element associated with antibiotic resistance, correlating with their resistance phenotype. This is consistent with data obtained from analysis of *V. cholerae* isolates from a previously published study (Kiiru, *et al.*, 2009). This data shows that antibiotic resistance is phenotypically expressed in both the environmental and clinical *V. cholerae* isolates. Interestingly, with the exception of two isolates, Kenyan clinical samples had an identical antibiotic resistance profile whereas the samples collected from environmental sources had varied resistance profiles, irrespective of where they clustered in the phylogenetic tree. We know that many of the resistance determinants like sulphamethoxazole, kanamycin,

chloramphenicol, streptomycin, tetracycline and trimethoprim could be associated with the SXT elements and that the SXT elements carry hot spots for recombinogenic activity within *V. cholerae*, providing a mechanism for more rapid evolution of resistance.

The assemblies of all non-O1 isolates that clustered outside the seventh pandemic lineage (Figure 3.6A) were also analysed for any novel regions inserted or deleted with respect to the El Tor reference genome of N16961. With three exceptions, all non-O1 isolates lacked well known virulence related elements such as VPI-1, VPI-2, VSP-1, VSP-2 and the cholera toxin phage CTX. The three exceptions were KNE056B_2, KNE17 and KNE150. KNE056B_2 and KNE17 possessed CTX and VPI-1, and KNE056B_2 also possessed VPI-2. Of note, KNE150 carried an R391-ICE inserted in the peptide release chain factor-3 gene (*prf*C-3), the site specific for the insertion of R391 family ICE element. A range of novel and previously identified islands was found among the highly diverse non-O1 genomes. All genomic islands found in these isolates were catalogued and are listed in Table 3.4. To identify the roles these genes may play, the phenotypic and genotypic characterization of these islands needs further work.

Genomic Islands in Non-O1/O139 Kenyan Environmental strains

Strain	i	ii	iii	iv	v	vi	vii	viii	ix	x	xi	xii	xiii	xiv	xv	xvi	xvii	xviii	xix	xx
KNE98	present	present	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE081A	present	present	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE102A	present	present	present	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE83	present	present	absent	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE59	present	absent	absent	present	absent	absent	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE85	absent	absent	absent	present	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE85C	present	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE96	absent	absent	absent	absent	absent	absent	absent	absent	present	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE096B	absent	absent	absent	absent	absent	absent	absent	absent	present	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE150	absent	present	present	present	absent	absent	absent	absent	absent	present	absent	present	present	absent	absent	absent	absent	absent	absent	absent
KNE10G	present	absent	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE45	absent	present	absent	present	absent	absent	present	absent	absent	absent	absent	present	absent	absent	absent	absent	absent	absent	absent	absent
KNE70	present	present	absent	present	absent	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE104C	present	present	present	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	present	absent	absent	absent	absent	absent	absent
KNE7	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE114	absent	present	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	absent	absent	absent	absent
KNE60	absent	absent	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	absent	absent	absent
KNE04C	absent	present	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE109B	absent	absent	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	absent	absent
KNE109A	absent	absent	present	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	absent	absent
KNE53	absent	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	absent
KNE18	absent	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	present	absent	absent
KNE81	absent	present	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	present
KNE056B_2	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE17	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent
KNE195	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	absent	absent
KNE083A	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	absent	present	absent	absent	absent	absent	absent	absent

absent
present

Key	Position in the chromosome
i	Insertion between VC0002-VC0003 (GI-15* as in Chun et al 2009)
ii	Insertion between VC1910-VC1911
iii	Insertion between VC2041-VC2042
iv	Insertion between VC2714-VC2715
v	Insertion between VC0422-VC0423
vi	Insertion between VC0031-VC0032
vii	Insertion between VC0768-VC0769
viii	Insertion between VC0978-VC0979
ix	Insertion between VC00487-VC0488
x	Insertion between VC00806-VC0807
xi	Insertion between VC1299-VC1300
xii	Insertion between VC0080-VC0081

Table 3.4: Table showing the presence or absence of novel genomic regions, listed across the top and described in the key, in the non-O1/O139 Kenyan isolates from the environment.

3.3.7 A novel *ctx*B gene in some Kenyan non-O1 environmental isolates

The *ctx*B gene type was analysed for each sequenced isolate. With the exception of KNE231 and KNE241 that harboured the *ctx*B-3b gene (Mutreja, *et al.*, 2011), all other Kenyan O1 El Tor isolates harboured the *ctx*B-3 toxin allele (Mutreja, *et al.*, 2011). During the whole genome sequence analysis, it was noticed that unusually the non-O1 environmental isolates, KNE056B_2 and KNE17, harboured an identical *ctx*B gene, which differed from the *ctx*B gene sequence of the El Tor reference N16961 by 14 SNPs. This *ctx*B gene represents an entirely novel sequence in any non-O1/O139 *V. cholerae* (Figure 3.8).

```
KNE17_ctxB          ATGATTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCATATGTACAT 60
KNE056B_2_ctxB      ATGATTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCATATGTACAT 60
J31W                ATGATTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCATATGTACAT 60
N16961_ctxB         ATGATTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCATATGCACAT 60
                    ************************************************************

KNE17_ctxB          GGAACACCACAAAATATTACTGATTTGTGTGCGGAATACAACAACACACAAATATATACG 120
KNE056B_2_ctxB      GGAACACCACAAAATATTACTGATTTGTGTGCGGAATACAACAACACACAAATATATACG 120
J31W                GGAACACCACAAAATATTACTGATTTGTGTGCGGAATACAACAACACACAAATATATACG 120
N16961_ctxB         GGAACACCTCAAAATATTACTGATTTGTGTGCAGAATACCACAACACACAAATATATACG 120
                    ******** ************************ ****** ********************

KNE17_ctxB          CTAAATGAAAAGATATTGTCGTATACAGAATCTCTAGCTGGAAAAAGAGAGATGGCTATC 180
KNE056B_2_ctxB      CTAAATGAAAAGATATTGTCGTATACAGAATCTCTAGCTGGAAAAAGAGAGATGGCTATC 180
J31W                CTAAATGAAAAGATATTGTCGTATACAGAATCTCTAGCTGGAAAAAGAGAGATGGCTATC 180
N16961_ctxB         CTAAATGATAAGATATTTTCGTATACAGAATCTCTAGCTGGAAAAAGAGAGATGGCTATC 180
                    ******** ******** ******************************************

KNE17_ctxB          ATTACTTTTAAGAATGGTGAAACTTTTCAAGTAGAAGTGCCAGGTAGTCAACATATAGAT 240
KNE056B_2_ctxB      ATTACTTTTAAGAATGGTGAAACTTTTCAAGTAGAAGTGCCAGGTAGTCAACATATAGAT 240
J31W                ATTACTTTTAAGAATGGTGAAACTTTTCAAGTAGAAGTGCCAGGTAGTCAACATATAGAT 240
N16961_ctxB         ATTACTTTTAAGAATGGTGCAATTTTTCAAGTAGAAGTACCAGGTAGTCAACATATAGAT 240
                    ******************** ** *************** ********************

KNE17_ctxB          TCACAAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGAATTGCATATCTTACTGAA 300
KNE056B_2_ctxB      TCACAAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGAATTGCATATCTTACTGAA 300
J31W                TCACAAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGCATATCTTACTGAA 300
N16961_ctxB         TCACAAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGCATATCTTACTGAA 300
                    ****************************************** ******************

KNE17_ctxB          GCTAAAGTTGAAAAGTTATGTGTATGGAACAATAAAACACCTAATGCGATTGCCGCAATT 360
KNE056B_2_ctxB      GCTAAAGTTGAAAAGTTATGTGTATGGAACAATAAAACACCTAATGCGATTGCCGCAATT 360
J31W                GCTAAAGTTGAAAAGTTATGTGTATGGAACAATAAAACACCTAATGCGATTGCCGCAATT 360
N16961_ctxB         GCTAAAGTCGAAAAGTTATGTGTATGGAATAATAAAACGCCTCATGCGATTGCCGCAATT 360
                    ******** ******************** ******** *** ******************

KNE17_ctxB          AGTATGGCAAATTAA 375
KNE056B_2_ctxB      AGTATGGCAAATTAA 375
J31W                AGTATGGCAAATTAA 375
N16961_ctxB         AGTATGGCAAATTAA 375
                    ***************
```

Figure 3.8: Multiple nucleotide sequence alignment performed using Clustal X 2.1 showing the *ctxB* sequences of KNE17, KNE056B, J31W and N16961 aligned. The base positions with * indicate a match and those with a gap indicate a mismatch.

Database searches revealed that the closest match to this novel *ctx*B gene was found in a non-O1 environmental isolate J31W reported from Argentina in 2009 (Genbank accession FJ748608), which differs by a single base pair from these Kenyan isolates at base position 282. The *ctx*B gene of J31W, in contrast, has the same sequence as the El Tor reference N16961 *ctx*B at this position. The alignment showing the *ctx*B genes of N16961, KNE056B_2, KNE17 and J31W is shown in Figure 3.8.

3.3.8 Whole genome phylogeny of Mexican *V. cholerae*

For the detailed understanding of Mexican *V. cholerae* population, whole genome sequencing was performed on a collection from the archives of the major Latin American cholera outbreak of 1991-1995 and the samples from sporadic cholera cases reported between 1991 and 2010. Samples were also collected from environmental sources such as river water, bottled water and food items for a more complete interpretation of *V. cholerae* diversity present in Mexico.

DNA extracted from 84 Mexican isolates was sequenced and this information was collated with data from 231 previously published clinical and environmental strains (Hasan, *et al.*, 2012; Hendriksen, *et al.*, 2011; Mutreja, *et al.*, 2011). All the Illumina paired end read data was mapped to the completed reference N16961 El Tor genome and a global whole genome phylogeny was constructed (Figure 3.9A). In the consensus tree, 49 isolates clustered with the global seventh pandemic El Tor clade and 6 clustered within the classical lineage. 10 other isolates shared ancestors with the US Gulf coast lineage (Figure 3.9A) but were grouped together on a separate branch approximately 10,000 SNPs away from the previously sequenced US Gulf coast isolates. Here, this cluster has been named “Mexican Local Endemic-1 (MLE-1)” lineage (Figure 3.9B). Figure 3.9B also shows a cluster of 4 isolates forming a new clade here referred to as “Mexican Local Endemic-2 (MLE-2)” lineage (Figure 3.9C). These lineages have been called local lineages because isolates of this genotype have not been described previously. 15 Mexican isolates formed diverse individual single isolate lineages, each more than ~54,000 SNPs away from the N16961 reference El Tor genome (Figure 3.9A).

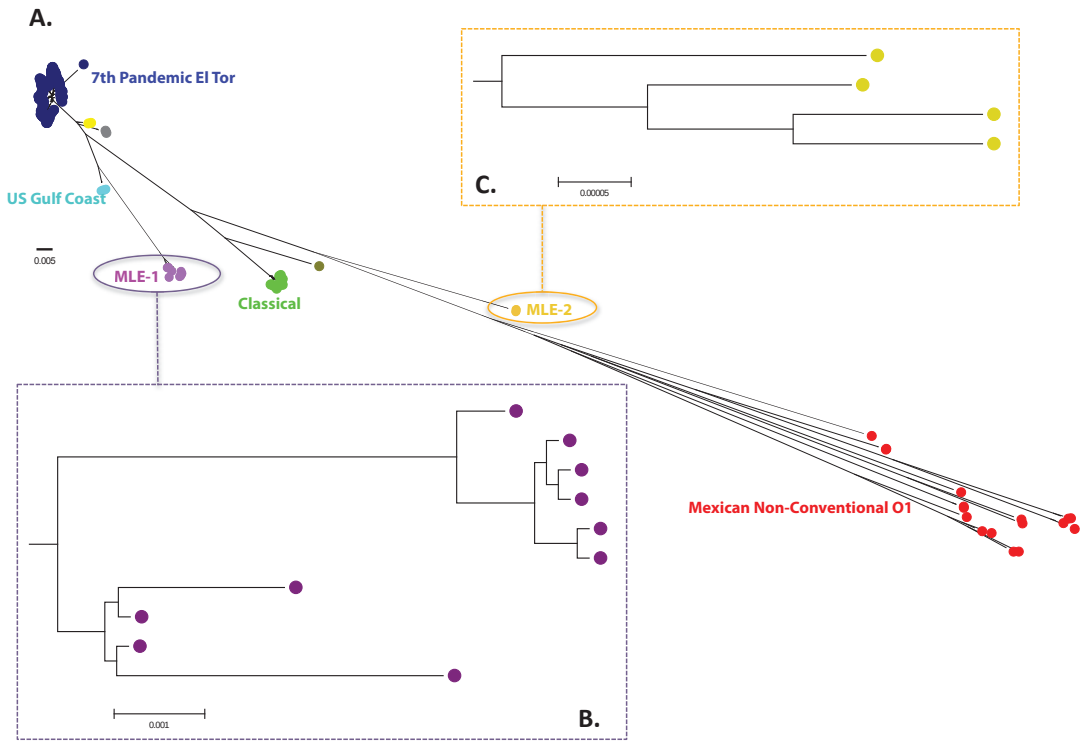


Figure 3.9: A) A Maximum Likelihood phylogenetic tree based on SNP variation. The major lineages are shown and MLE-1 and MLE-2 lineages are circled. In red are the non-conventional O1 isolates from Mexico. B) A mid-point rooted maximum-likelihood phylogenetic tree of the MLE-1 isolates. C) A mid-point rooted maximum-likelihood phylogenetic tree of the MLE-1 isolates. All the scales are given as the number of substitutions per variable site.

As shown in Figure 3.10, of the 49 Mexican isolates that clustered in the seventh pandemic El Tor lineage, 32 clustered within the WASA-1 cluster (section 2.3.9) in wave-1 alongside other Latin American isolates from Argentina, Bolivia, Colombia and Peru. Interestingly, the other 17 clustered in wave-2 of the cholera seventh pandemic lineage. Amongst them was also an O139 serogroup isolate EM-0892, which clustered within the O139 lineage in wave-2.

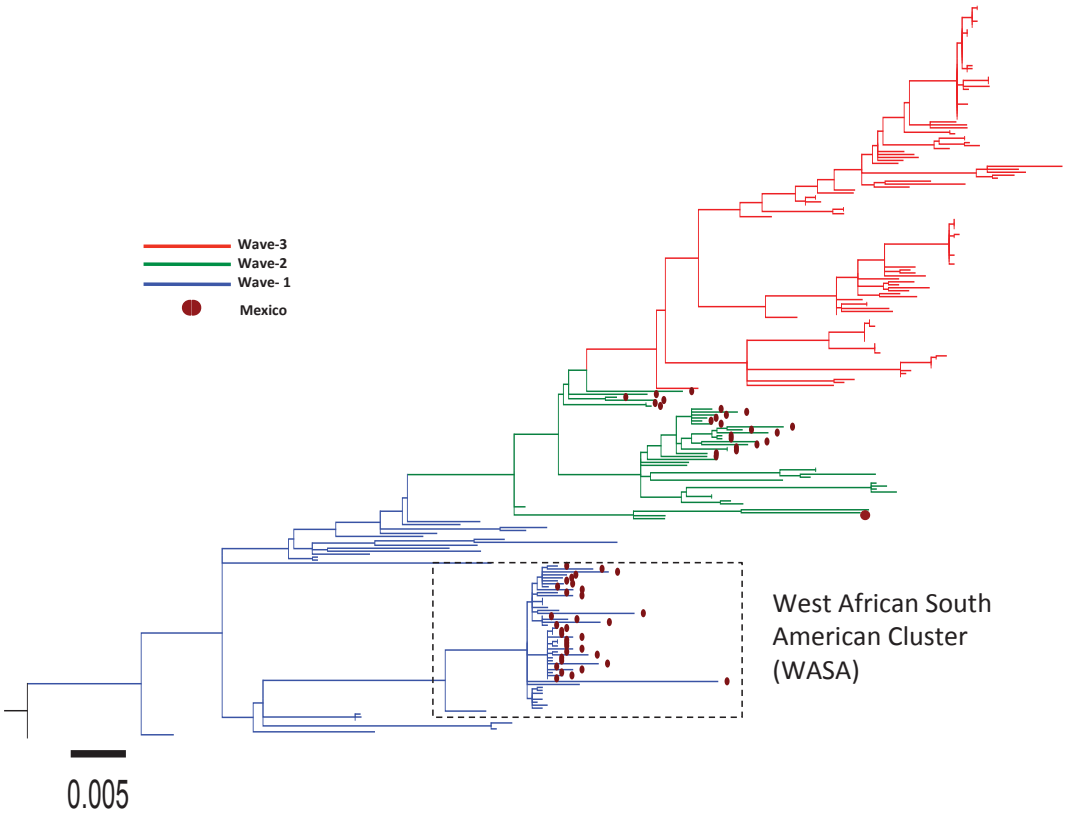


Figure 3.10: Maximum Likelihood phylogenetic tree of the seventh pandemic lineage including the Mexican seventh pandemic isolates from our collection marked as brown circles in the tree. The phylogeny is based on SNP variation after excluding the high density SNPs and likely recombination events. Completed reference genome of N16961 was used as the reference against which to map the genomes of all the isolates and a pre-seventh pandemic isolate, M66, was used as an out-group to root the tree. The scale is given as the number of substitutions per variable site.

The presence of Mexican isolates in both wave-1 and wave-2 suggests that there were at least two separate introductions of cholera in Mexico during the Latin American cholera epidemic period. Within wave-1, all of the Mexican isolates fell within the lineage characterized by the presence of WASA-1 (explained in detail in section 2.3.9). Isolates in WASA-1 lineage of wave-1 are predicted to be an introduction from south-Asia possibly *via* Portuguese speaking West African countries. Clustering of the wave-2 strains within wave-2 represents the other introduction, which could have been either directly from South Asia or *via* Africa.

3.3.9 Genomic islands and new markers in the Mexican *V. cholerae* genomes

To look for any genomic markers that could differentiate the different Mexican clades, all the 84 genomes were assembled and manually compared against the completed reference N16961 genome. All the isolates that clustered within the seventh pandemic lineage had complete O1 antigen cluster and possessed Vibrio pandemicity islands VPI-1 and 2 and Vibrio seventh pandemic marker islands VSP-1 and 2. However, to our surprise, all the Mexican seventh pandemic isolates including those isolated in 2010 were genotypically SXT negative.

The findings of manual assembly comparisons agreed with the previous findings (see section 2.3.9) for wave-1 Mexican isolates as all these had the WASA-1 phage inserted between VC1494 and VC1495 (Mutreja, *et al.*, 2011). Also, the genes VC0512-VC0516 of VSP-2 island were replaced by homologous recombination as reported previously (Mutreja, *et al.*, 2011). All, but isolate 8338 (isolated from a clinical sample from Tabasco in 1991), of the isolates in wave-1 harboured a CTX phage. However, unusually, despite lacking the CTX phage, strain 8338 clustered in the WASA lineage with the wave-1 South American isolates based on the whole genome SNPs. This finding was phenotypically confirmed by our collaborators in Mexico who used ELISA for demonstrating the absence of cholera toxin production by this isolate.

All the wave-2 Mexican isolates, in contrast to wave-1 isolates, lacked the WASA-1 phage but were found to be carrying the previously discovered GI-15 (Chun, *et al.*, 2009; Mutreja, *et al.*, 2011), a kappa prophage inserted between the CDSs VC0002 and VC0003. The CTX phage was present in every wave-2 Mexican isolate and the sequence of *ctx*B gene was of CTX-2 type as opposed to the CTX-1 type present in the wave-1 Mexican isolates. One Mexican isolate, EM_0892, clustered within the O139 lineage and therefore the CTX phage harboured a *ctx*B type distinct from both CTX-1 and CTX-2 (Figure 3.11). Homologous recombination of the O-antigen cluster from a source outside the seventh pandemic tree was clearly noticeable in the genome of EM_0892. All the wave-2 isolates, except 33297, in wave-1 and wave-2 agreed genotypically and phenotypically on their O1 serogroup characterization. Strain

33297 was phenotypically serotyped as O36, but it did not show any replacement of the O1-antigen cluster genes.

```
ctxB_CTX_2          ATGATTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCATATGCACAT 60
ctxB_CTX_O139       ATGATTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCATATGCACAT 60
ctxB_CTX_1          ATGATTAAATTAAAATTTGGTGTTTTTTTTACAGTTTTACTATCTTCAGCATATGCACAT 60
                    ************************************************************

ctxB_CTX_2          GGAACACCTCAAAATATTACTGATTTGTGTGCAGAATACCACAACACACAAATACATACG 120
ctxB_CTX_O139       GGAACACCTCAAAATATTACTGATTTGTGTGCAGAATACCACAACACACAAATATATACG 120
ctxB_CTX_1          GGAACACCTCAAAATATTACTGATTTGTGTGCAGAATACCACAACACACAAATATATACG 120
                    ************************************************** *****

ctxB_CTX_2          CTAAATGATAAGATATTTTCGTATACAGAATCTCTAGCTGGAAAAAGAGAGATGGCTATC 180
ctxB_CTX_O139       CTAAATGATAAGATATTTTCGTATACAGAATCTCTAGCTGGAAAAAGAGAGATGGCTATC 180
ctxB_CTX_1          CTAAATGATAAGATATTTTCGTATACAGAATCTCTAGCTGGAAAAAGAGAGATGGCTATC 180
                    ************************************************************

ctxB_CTX_2          ATTACTTTTAAGAATGGTGCAACTTTTCAAGTAGAAGTACCAGGTAGTCAACATATAGAT 240
ctxB_CTX_O139       ATTACTTTTAAGAATGGTGCAACTTTTCAAGTAGAAGTACCAGGTAGTCAACATATAGAT 240
ctxB_CTX_1          ATTACTTTTAAGAATGGTGCAATTTTTCAAGTAGAAGTACCAGGTAGTCAACATATAGAT 240
                    ********************** ***********************************

ctxB_CTX_2          TCACAAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGCATATCTTACTGAA 300
ctxB_CTX_O139       TCACAAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGCATATCTTACTGAA 300
ctxB_CTX_1          TCACAAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGCATATCTTACTGAA 300
                    ************************************************************

ctxB_CTX_2          GCTAAAGTCGAAAAGTTATGTGTATGGAATAATAAAACGCCTCATGCGATTGCCGCAATT 360
ctxB_CTX_O139       GCTAAAGTCGAAAAGTTATGTGTATGGAATAATAAAACGCCTCATGCGATTGCCGCAATT 360
ctxB_CTX_1          GCTAAAGTCGAAAAGTTATGTGTATGGAATAATAAAACGCCTCATGCGATTGCCGCAATT 360
                    ************************************************************

ctxB_CTX_2          AGTATGGCAAATTAA 375
ctxB_CTX_O139       AGTATGGCAAATTAA 375
ctxB_CTX_1          AGTATGGCAAATTAA 375
                    ***************
```

Figure 3.11: Multiple nucleotide sequence alignment performed using Clustal X 2.1 showing the *ctxB* sequences of representatives of wave-2 O1, wave-2 O139 and wave-1 O1 Mexican isolates. The base positions marked with * indicate a match and those with a gap indicate a mismatch.

Corresponding to their position in the global tree, the isolates in the non-pandemic lineages MLE-1 and MLE-2 were very diverse at the genome level. However, despite being more than ~16,000 and ~25,000 SNPs different from the seventh pandemic lineage, respectively, all the isolates harbored loci similar to other well-characterized O1-LPS determinants in known O1 positive *V. cholerae*. However, MLE-1 isolates, similar to the US Gulf coast strains, possessed VPI-1 and 2 but lacked VSP-1 and 2. While the presence of CTX phage has been reported in some of the US Gulf coast strains previously (Chun, *et al.*, 2009), all MLE-1 isolates lacked the CTX phage genes except the isolate 3056, which lacked the *ctx*AB genes responsible for the cholera toxin production but had other core CTX genes (*zot*, *ace*, *orfU*, *cep*). Interestingly, a gene likely encoding a protein with a mucinase activity domain was

inserted at a specific site (between CDSs VC1587-VC1588) in the genomes of all the MLE-1 isolates. The insertion of this gene is unique to this lineage and could play an important role in modifying the host cell surfaces.

In contrast to MLE_1 isolates, all the isolates of the MLE-2 lineage carried VPI-1 island but lacked VPI-2, VSP-1 and VSP-2. Moreover, the CTX phage was completely absent from all MLE-2 isolates, except 2714 and 2710, which like 3056 of MLE-1, had core CTX genes but lacked the toxin producing genes *ctx*AB. Two prophages, inserted at two specific insertion sites (VC0217-VC0218 and VC2041-VC2042), were found to be unique to the MLE-2 lineage and have been named here as MLE-2a and MLE-2b phage, respectively. A NCBI database search of these found that DNA sequences from both these phages showed significant homology to a single contig in a whole genome shotgun sequence of *V. cholerae* strain CP1037(10) (accession number NZ_JH942263), which interestingly was isolated in Mexico in 2003. To determine how this isolate relates phylogenetically to MLE-2, the genome sequence of CP1037(10) was analysed and it clustered within the MLE-2 lineage (Figure 3.12) and was only 75 SNPs different from the MLE-2 islate 1146.

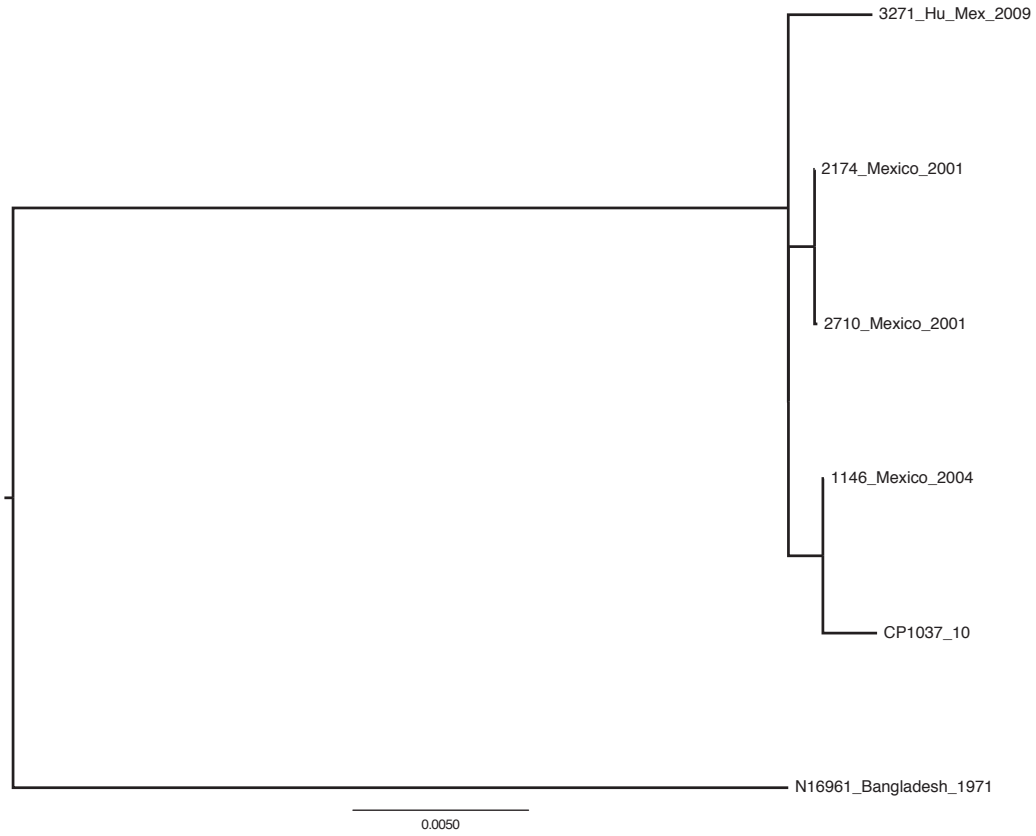


Figure 3.12: Maximum Likelihood phylogeny showing the four MLE-2 isolates and CP1037(10) from NCBI database search results. N16961 was used as the reference to map the sequence data. The scale is given as the number of substitutions per variable site.

Analysis of the diverse Mexican non-conventional O1 isolates (coloured red in Figure 3.9A) showed that some possessed regions of differences that were either novel or had been described previously (Chun, *et al.*, 2009; Mutreja, *et al.*, 2011). These isolates fell outside of the seventh pandemic cluster (Chun, *et al.*, 2009) and matched the genomic features of the previously reported strains. Although they possess a genomic backbone more than 54,000 SNPs different from the seventh pandemic lineage strains, they also harboured a similar O1 serogroup antigen gene cluster (see section 2.3.1). The only exception was 2806, which had the O1-antigen cluster replaced by other genes. This agrees with the phenotypic serotyping, which showed 2806 to be of O14 serogroup. Interestingly, 4 of the non-conventional O1 isolates (2709, 2370, 1474 and 1148) had the R391 family SXT ICE element inserted in their *prf*C-3 gene.

3.4 Conclusion and lessons from the regional case studies

In the first detailed study of the molecular epidemiology of *V. cholerae* from Pakistan, whole genome sequencing and SNP-based phylogenetic analyses was able to provide some epidemiological answers about the spread of cholera in Pakistan during the floods of 2010. The geographic distribution of the isolates in PSC-1 and PSC-2 was particularly revealing as isolates from PSC-1 were largely limited to the non-flood affected coastal city of Karachi and only one PSC-1 isolate was from the nearby city of Hyderabad, whereas isolates from PSC-2 were from inland flood and non-flood affected areas countrywide (Figure 3.1, 2). A few sporadic cases alongside the two sub-clades suggest that during the floods there were two or possibly three routes of cholera spread in Pakistan: one along the course of the Indus river, a second from the Arabian sea and the third possible route could be with infected travellers or contaminated food. The position of the PSC-1 and PSC-2 isolates on the global phylogenetic tree of *V. cholerae* O1 places them close to *V. cholerae* from India

isolated in 2006 and 2007 and isolates from Bangladesh and India from 2004 and 2005 respectively (Mutreja, *et al.*, 2011) (Figure 3.2). The phylogeny of the two sub-clades unequivocally shows that PSC-1 and PSC-2 have evolved from two different recent ancestors. Thus, during the floods at least two sub-clades of *V. cholerae* co-existed in Pakistan with different patterns of spread indicating an interesting epidemic within an epidemic scenario.

In the Kenyan surveillance study, other than the two sub-clades (KSC-1 and KSC-2), the presence of environmental *V. cholerae* isolates phylogenetically distinct from the main El Tor lineage is particularly noteworthy. And equally important is the finding of the El Tor lineage isolates that could be sourced from the environment. While the diverse strains outside the seventh pandemic may be associated with diarrheal diseases distinct from cholera in the respective local communities, the isolation of the seventh pandemic lineage strains from the environment highlights the possibility of contamination of the environmental resources by the isolates that have outbreak causing capability. The existence of non-epidemic lineage isolates with the clinically important *V. cholerae* El Tor isolates in the environment presents increased chances of genetic recombination and the exchange of antibiotic resistance determinants between these phylogenetically distinct populations. The horizontal transfer of toxin genes, crucial O-antigen genes and pandemicity islands to the non-epidemic lineage strains could give lead to an increased cholera burden. Similarly, the transfer of antibiotic resistance gene cassettes from the non-epidemic pool of strains that are greatly exposed to the environmental stress and challenges to the epidemic lineage strains could severely cut the spectrum of antibiotics available for treating cholera cases in hospitals. Investigation of the environmental and food sources alongside the clinical samples is recommended for inclusion in future studies to obtain the full breadth of the *V. cholerae* populations circulating in any particular region.

In Chapter 2, only wave-1 isolates of *V. cholerae* were identified in the limited number of Latin American isolates included in this study. However, this detailed study on Mexican isolates highlighted that not only wave-1 but also wave-2 isolates entered Mexico during the cholera outbreaks of 1991-1996. According to their position in the phylogeny (Figure 3.10), there must have been at least two separate introductions of isolates into Mexico all originating from the nodes associated with

the south-Asian *V. cholerae*, but whether this pattern applies to the whole South American continent remains to be checked.

Further, the Mexican isolates that clustered in the wave-1 WASA cluster spanned the years 1991 to 2010, and the wave-2 isolates spanned 1991-2002. This suggests that while the wave-1 *V. cholerae* have persisted in Mexico since their entry in the country as part of the general Latin American pandemic, wave-2 Mexican isolates entered independently. This is the first time that we have identified this particular phylogenetic lineage of *V. cholerae* O1 El Tor persisting at this time within a particular endemic area. This finding is perhaps even more intriguing because all the Mexican isolates from both wave-1 (including those recently isolated from 2010) and wave-2 (except the O139 isolate EM-0892) lack the R391 family SXT ICE element in their genomes. Since SXT is found in other seventh pandemic *V. cholerae* isolated recently elsewhere in the world, it would be worth investigating whether the seventh pandemic Mexican *V. cholerae* O1 strains harbor the resistance determinants, normally associated with SXT elements, on their superintegron.

4. The genetic basis of serotype variation in *V. cholerae* sampled during a clinical trial in Kolkata, India

NOTE: All the clinical trial isolates were collected by our collaborators based in Kolkata, India. The isolates were collected from 2003-2010 but only 2003 to 2007 were made available and the information about the patients whether they were vaccinated or not was not available to include in the analysis. The DNA was sent to the Sanger Institute for sequencing by the sequencing pipeline teams and raw short read data was made available for the analyses. The work explained in this chapter details the phylogenetic analyses, which was done by me and therefore forms a part of my PhD thesis.

4.1 Introduction

V. cholerae LPS is composed of an antigenically variable O-antigen polysaccharide (PS) and a core-PS, including lipid A, that exhibits relatively limited variation. *V. cholerae* expressing antigenically distinct O-antigen PS can be classed using antibodies to O-antigen into different serogroups. Using this approach, more than 200 *V. cholerae* serogroups have been identified to date, although only O1 and O139 strains can cause epidemic cholera. In addition to biotyping (classical and El Tor), *V. cholerae* O1 and O139 isolates can be further divided into serotype Ogawa or Inaba (Figure 4.1) based on further antigenic properties of the O-PS.

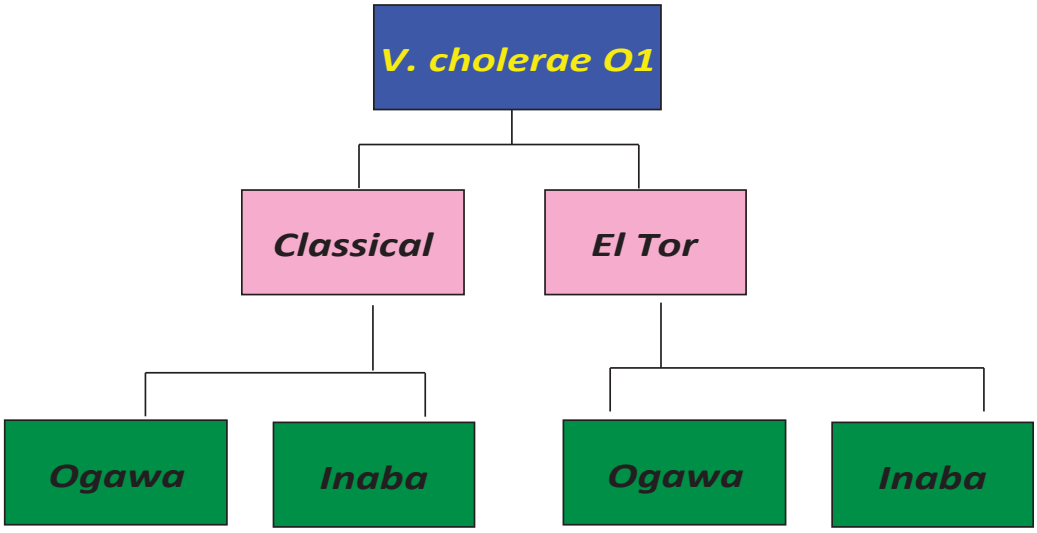


Figure 4.1: Classification of O1 *V. cholerae* into biotypes and serotypes. *V. cholerae* O1 has two biotypes Classical and El Tor and both biotypes can have two serotypes Inaba and Ogawa.

The presence or absence of a single methyl group on the terminal sugar of the O-PS can change the serotypic response of a *V. cholerae* isolate from Ogawa to Inaba. The proteins that drive the synthesis of the O1-antigen of *V. cholerae* are predominantly encoded by the '*wbe* region' of the genome, which is between 16-19 kb, varying between isolates. The *wbe*T gene, a methyl transferase, in the *wbe* operon encodes an enzyme responsible for the methylation of terminal 4-N-tetronylated-D-perosaminyl group (Chatterjee and Chaudhuri, 2003) on O-PS and the Inaba-Ogawa distinction has been correlated to the alteration in this gene (Stroeher, *et al.*, 1992). If *wbe*T is expressed in its wild type form, the resultant phenotype is Ogawa but if the gene is missing or does not drive the expression of the functional enzyme, the resultant phenotype is Inaba. Indeed, the introduction of a complete *wbe*T gene from an Ogawa into an Inaba isolate can mediate conversion to Ogawa (Stroeher, *et al.*, 1992). Interestingly, spontaneous conversion from Inaba to Ogawa occurs at a much lower relative frequency compared to Ogawa to Inaba conversions. One possible reason for this could be that while Ogawa to Inaba conversion would just require a mutation in *wbe*T, the Inaba to Ogawa conversion would need the parsimoniously much less likely event of mutation correction (for example by a recombination event). This fixation event is likely to be comparatively rare through natural evolutionary events

but could be favoured in a population if there is a strong selection for Ogawa in any specific environment.

The work described in this chapter is divided into two parts. First, the sequences of seventh pandemic O1 El Tor *V. cholerae* were screened informatically to identify mutations in *wbe*T that might influence the expression of the Ogawa and Inaba epitopes. Fortunately, for most of the sequenced isolates, there was phenotypic information available for serotype. The aims were to (a) identify common mutations within *wbe*T associated with the Ogawa to Inaba conversion; (b) determine if there were any examples where a phenotypic change in serotype did not correspond to a clear mutation in *wbe*T; (c) identify any likely examples of conversion of Inaba to Ogawa i.e. any correction of *wbe*T to wild type based on phylogenetic position; (d) identify isolates harbouring the wild type *wbe*T allele that are phenotypically Inaba. To achieve this, the *wbe*T gene of sequenced seventh pandemic *V. cholerae* were directly compared to the wild type El Tor Ogawa *wbe*T allele.

Evidence has accumulated that exposure of an individual to a *V. cholerae* Inaba challenge can provide some subsequent protection against both Inaba and Ogawa infection, whereas an Ogawa challenge only protects against an Ogawa infection (Longini, *et al.*, 2002). Consequently, the genetic basis of the Inaba-Ogawa variation was determined in a sample set of *V. cholerae* collected in a vaccine trial undertaken in Kolkata, India (Sur, *et al.*, 2009). The design of cholera vaccines has to take consideration of the serotype of the O1 *V. cholerae* used to manufacture the vaccine, particularly in the case of whole cell formulations. Consequently whole cell-based cholera vaccines currently on the market harbour killed *V. cholerae* cells as a mixture of both Inaba and Ogawa serotype strains (section 1.2.6), to get better overall protection. Thus, any knowledge about the serotypic composition of *V. cholerae* in a particular geographical region and the mechanisms by which serotypic switching might occur would be of practical value.

4.2 Results and discussion

4.2.1 *wbeT* sequence analysis

The genome sequences of each of the 1002 *V. cholerae* in the expanded seventh pandemic phylogeny (chapter 5) was assembled and the *wbe*T gene was determined to be of sufficient quality to analyse in detail in 777 of these sequences. The remaining 225 assemblies either had a contig break in *wbe*T or had poor coverage in this region. Of the 777 analysed *wbe*T gene sequences 244 were found to possess likely mutations in *wbe*T compared to the wild type allele. The types of mutations identified are shown as percentages in Figure 4.2. Non-synonymous single amino acid changes were the most frequent mutations in *wbe*T followed by frame shift mutations and insertions. The formation of stop codons, likely linked to premature termination, were more common than deletions.

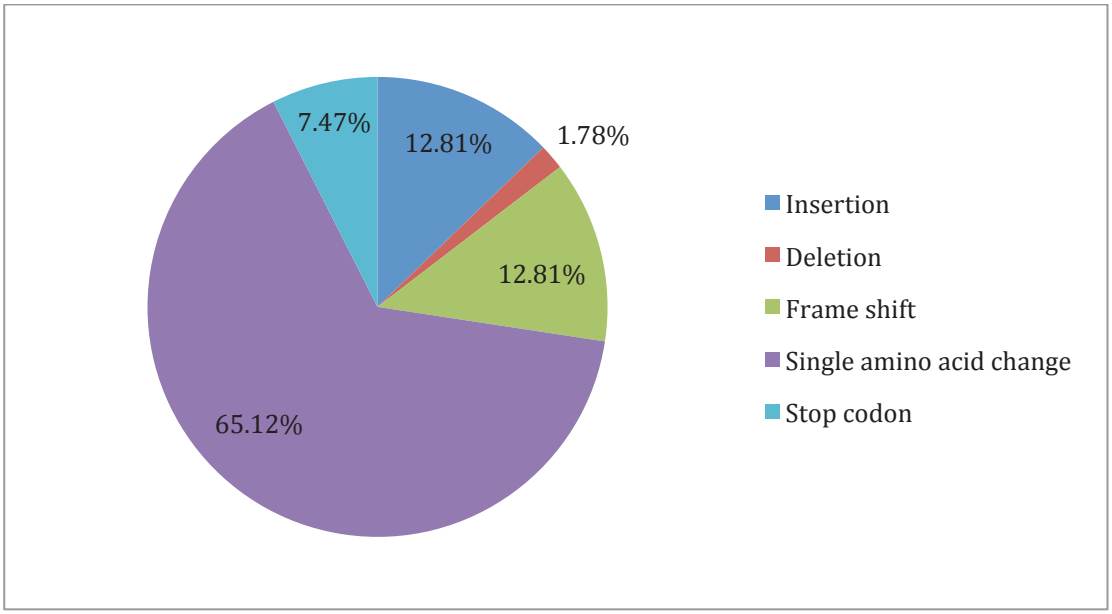


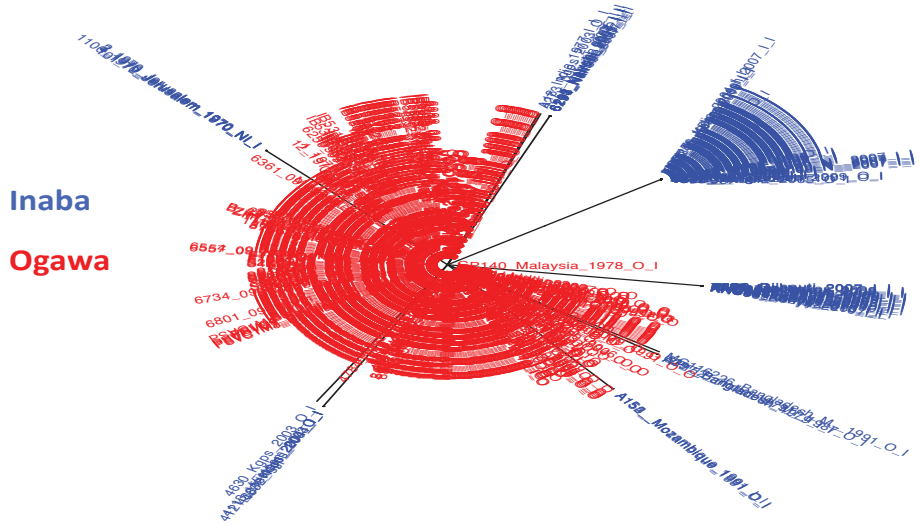
Figure 4.2: Pie-chart showing the percentage distribution of different types of mutations in the *wbeT* gene of 244 *V. cholerae* O1 El Tor sequences.

After identifying *V. cholerae* harbouring mutations in *wbe*T, a maximum likelihood phylogeny was constructed based on the *wbe*T gene alignment of all 777 analyzable sequences. This phylogenetic tree was used as a platform to visualize how well the phenotypic and genotypic characterization of serotype matched. As the reference *V. cholerae* N16961 was used to build the phylogeny and all the isolates with a wild type *wbe*T allele clustered together as one group whereas isolates with mutations formed

several small groups depending on the type of mutation (Figure 4.3A).

Since any inactivating mutation in the wild type Ogawa *wbe*T gene should result in a dysfunctional methyl transferase and therefore an Inaba conversion, serological phenotype should superimpose on this tree. However, although this was predominantly the case, there there were some phenotypically Ogawa isolates that mapped within clades with mutant *wbe*T and there were some phenotypically Inaba isolates located within clades of wild type *wbe*T (Figure 4.3B).

A.



B.

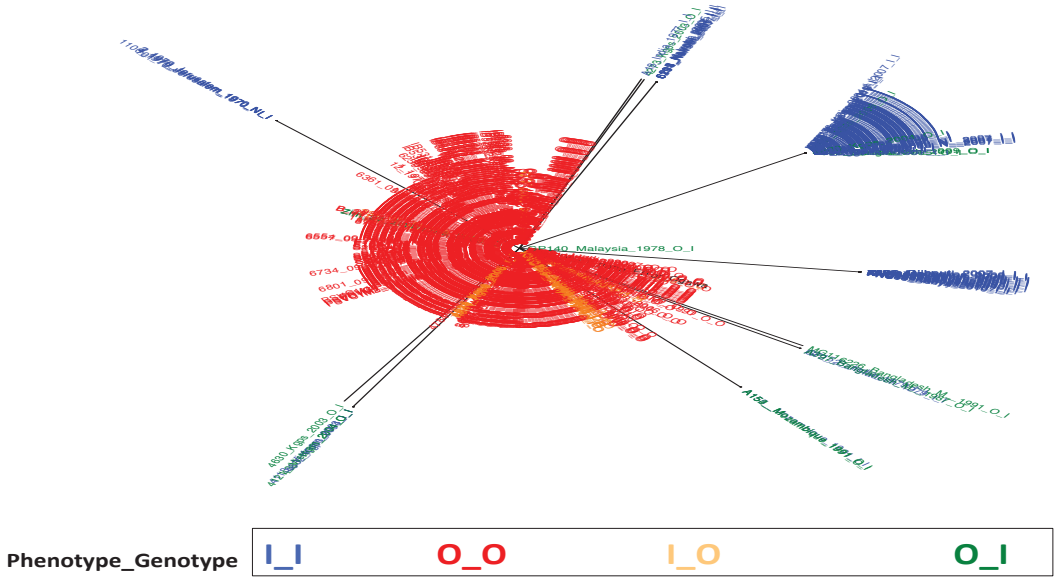


Figure 4.3: A) Maximum likelihood phylogenetic tree coloured as the Ogawa-Inaba strain arrangement based on genotype; B) The same phylogenetic tree showing how correlated phenotypic and genotypic data. I = Inaba; O = Ogawa.

Approximately ~90% of the isolates mapped onto the tree in accordance with both their genotype and phenotype, i.e. if they did not harbour any obvious mutation in the *wbe*T gene and were phenotypically Ogawa or if they did harbour a mutation in *wbe*T and they were phenotypically Inaba. However, ~10 % of the isolates did have a phenotype-genotype mismatch (Figure 4.3B, 4).

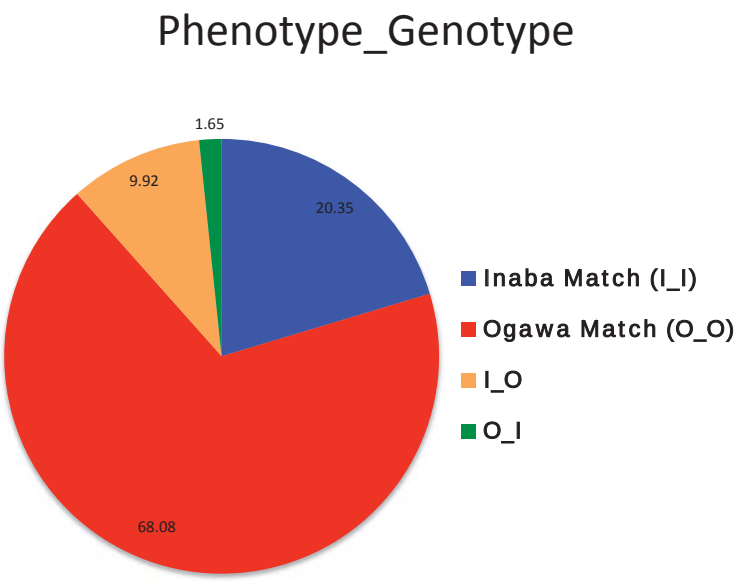


Figure 4.4: Pie-chart showing the percentage match and mismatch between the phenotypic and expected genotypic serotype. I = Inaba; O = Ogawa.

How can mismatches be explained? There were some mismatches where serotypically Inaba isolates harboured a wild type Ogawa *wbe*T allele. Such a combination of genotype and phenotype would be anticipated if the isolates harboured a mutation outside *wbeT*, which prevented the methyl transferase responsible for the methylation of terminal sugar from reaching the target, either through lack of expression or altered intracellular targeting. For example, a mutation could be in regions of the genome such as the promoter or a regulatory gene that influenced mRNA production or even

translation. Alternatively, the isolates could have been incorrectly serotyped, a relatively common phenomenon in routine serotyping laboratories (our unpublished observations).

There were a few isolates that harboured a mutated *wbe*T but were still phenotypically serotyped as Ogawa. This could be explained if such a mutation(s) did not have any effect on the expression of a functional enzyme or compensatory mutations were present elsewhere in the genome. Of course, mistakes in serotyping could also be an explanation here as well. Further work in the field will be required to investigate these possibilities in more detail.

4.2.2 Mapping *V. cholerae* from a vaccine trial performed in Kolkata to the global El Tor phylogeny

DNA from 405 *V. cholerae* O1 El Tor isolates collected from 405 cholera patients during a phase III field trail performed in Kolkata on a Shancol cholera vaccine were sequenced and analysed. This collection spanned five years, 2003-2007, with 2006 being the vaccination year (Sur, *et al.*, 2009). When these 405 *V. cholerae* were arranged in temporal order of their date of isolation and their phenotypically determined serotype was superimposed, an interesting pattern became apparent (Figure 4.5). In 2003 and 2004 Ogawa *V. cholerae* O1 El Tor isolates dominated but there was a relative increase in the number of isolates phenotypically Inaba in 2005, such that Ogawa isolates became the minority. Inaba isolates remained the dominant serotype in 2005 and 2006 but then there was an apparent decline in their population and Ogawa isolates re-emerged in 2007.

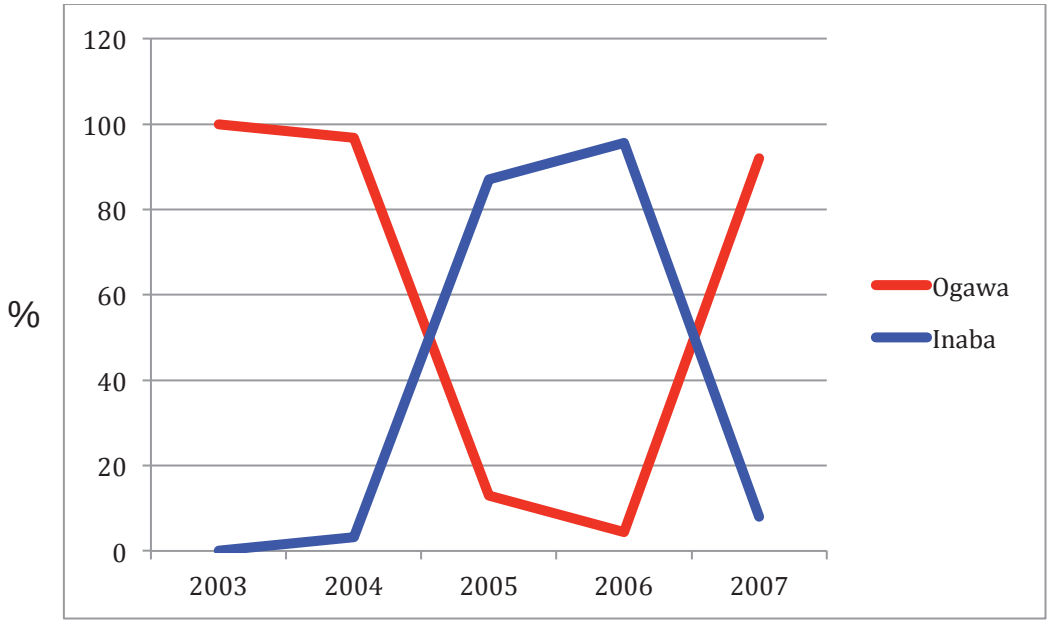


Figure 4.5: Distribution of Inaba and Ogawa serotype isolates during a clinical trial study in Kolkata, India. The data for isolates collected between 2007-2010 was not available.

Subsequently, a maximum likelihood phylogenetic tree was constructed to ascertain the position of these *V. cholerae* within the global seventh pandemic framework (not shown). All the isolates clustered in wave-3 of the seventh pandemic lineage and were distributed in both sub-clades 3a and 3b (explained in section 5.2.2.1). A further phylogenetic tree was constructed that included only the clinical trial isolates (Figure 4.6), using N16961 O1 El Tor as both reference and an outgroup, with the aim of obtaining a clearer understanding of their temporal and serotype distribution in a phylogenetic context. Interestingly, a clear temporal clustering of isolates was observed and the pattern of serotypic change from year to year corresponded well with the temporal clades in the tree.

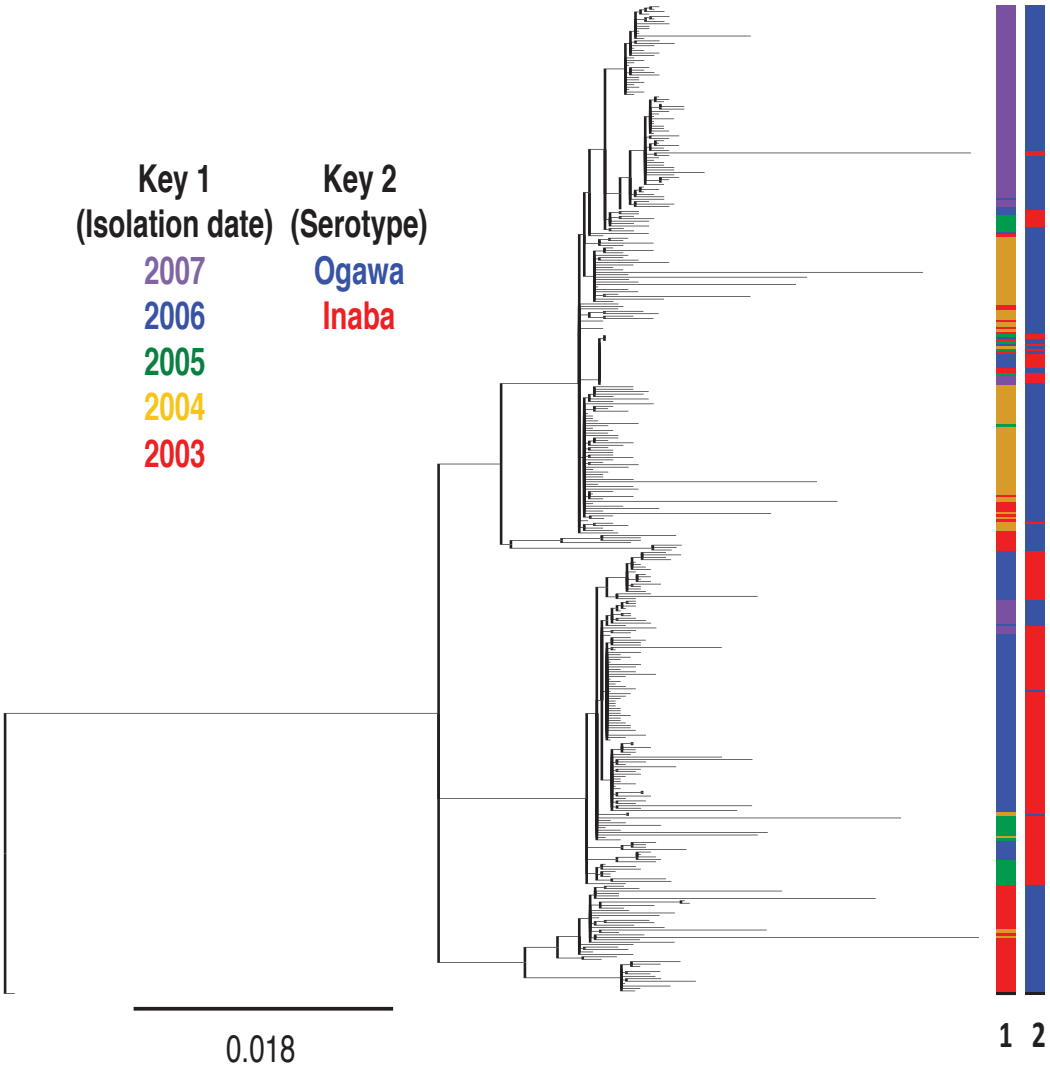


Figure 4.6: Maximum likelihood phylogeny of *V. cholerae* from the Kolkata vaccine trial with N16961 El Tor as both reference and root. Key 1 shows the year of isolation and Key 2 shows the serotype identified by antisera agglutination. The scale is given as substitution per variable site. The serotype switch from year to year is clear and there was visible temporal and serotypic correlation with the phylogeny.

4.3 Lessons learned and questions arising from this study

This study highlighted some aspects of the complexity of serotype variation in O1 *V. cholerae*. With the phenotypic agglutination test results and the genotypically predicted serotype matching for ~90% of the *V. cholerae* isolates, *wbeT* sequence appears to be a reasonably good marker for the genotypic classification of serotypes of *V. cholerae* O1. However, there were a few exceptions with a mismatch between

the phenotypically and genotypically determined serotype. *wbeT*, encoding the methyl transferase responsible for the methylation of the terminal sugar on O1-PS and consequently the Ogawa serotype of *V. cholerae* O1, harboured some mutations that were predicted to be null but did not apparently impact on serotype. There were also *V. cholerae* that harboured the wild type *wbeT* allele that reported as agglutinating with Inaba antisera. These data indicates that there may be mutations, for example polar mutations outside this gene, which influence the expression or the functionality of the final methyl transferase product.

Also, the data presented here prompts a number of potentially interesting questions including:

- Is a particular *wbeT* Inaba genotype prevalent in regional collections such as those from Kenya, Pakistan and Mexico ?
- Is a particular *wbeT* Inaba genotype found more frequently in the vaccine trial dataset ?
- Is there a particular mutated *wbeT* type which more fequently reverts back to the wild type Ogawa *wbeT*?
- Can we use similar data sets to predict if a selective pressure is operating in the field on the Inaba to Ogawa switch ?
- Is the change in serotype of *V. cholerae* population a switch driven by selection or is it simple strain replacement ?
- Is the mutation rate in *wbeT* gene the same as other genes of the wbe operon ?
- How different is the mutation rate in *wbeT* to the natural evolution rate of 3.3 SNPs/year in the seventh pandemic *V. cholerae* genomic backbone ?

Although these are all interesting questions there is insufficient data within the sample sets analysed within this study and further work must be planned to address these issues.

5. Expanded analysis of the seventh pandemic *V. cholerae* lineage and design of a PCR based SNP typing assays

NOTE: All the isolates were collected by our global collaboration partners. The DNA was sent to the Sanger Institute for sequencing by the sequencing pipeline teams and raw short read data was made available for the analysis. The work explained in this chapter was done by me and details the global phylogenetic analysis, which expands on the previous global analysis.

5.1 Introduction

Whole genome sequence analysis of global *V. cholerae* isolates revealed that seventh pandemic *V. cholerae* O1 El Tor has evolved from a single source population independently of the classical biotype (Mutreja, *et al.*, 2011). The pattern of genome wide SNPs in these seventh pandemic *V. cholerae*, as detailed in chapter 2, prove that the current pandemic is continuously evolving in a clock-like manner and is spreading in independent but overlapping waves from the source population. The global spread of this population, radiating in waves from the endemic Bay of Bengal region (Mutreja, *et al.*, 2011), is likely aided by modern travel behavior and the expanding food chain. Once *V. cholerae* has reached a previously non-endemic region, these isolates can cause local, regional and national level outbreaks. From public health perspective, it becomes imperative that the roots of any such spread are quickly and robustly traced back and appropriate actions taken.

As *V. cholerae* El Tor isolates are monophyletic, options for the development of classical typing approaches have been relatively limited. One of the approaches of choice was based on differences in sequence within the genes encoding cholera toxin harboured by the CTX phages. However, since CTX are mobile genetic elements, they do not evolve at the same rate or even within the same lineage as the genomic backbone and therefore cannot be trusted for true phylogenetic inference (see section 1.2.11). Molecular approaches, including MLST have little discriminatory power and are also of limited value. Although PCR based approaches based on changes in short repeat elements such as MLVA can detect diversity (Mohamed, *et al.*, 2012) they

have limited phylogenetic value. Recently, PFGE typing has become a preferred technique in some national reference centers and public health laboratories. Indeed, it is the current gold standard technique for the sub-typing of *V. cholerae* outbreak isolates. However, again PFGE provides limited phylogenetic information but rather reports on changes in phage elements, restriction sites and broader genome rearrangements. It is also a technique that generates data that is open to interpretation and is difficult to transfer between laboratories.

The work presented here has shown that the analysis of genome wide SNPs can discriminate between highly clonal lineages of *V. cholerae* and provide a phylogenetic context. Thus, it is perhaps a definitive approach to type both closely and distantly related *V. cholerae* isolates. The work presented here also provides evidence for strict clock-like evolution and limited recombination within *V. cholerae* populations, thus, SNP based approaches can provide a much needed level of resolution to monitor the real time spread of *V. cholerae*, providing a global context. Therefore, in this study, a SNP based typing scheme was designed, making practical use of high-resolution whole genome data obtained from global and regional seventh pandemic studies. This SNP genotyping assay could support real time surveillance of the on going cholera pandemic, potentially providing relatively quick and accurate monitoring of future cholera outbreaks in the context of previous ones.

1757 SNPs were detected during the initial sequence analysis of *V. cholerae* within the seventh pandemic lineage L2 (section 2.3.2) and these were used to build the phylogenetic tree (Figure 5.3 in chapter 2). 27 of these SNPs were selected from the stable regions of the *V. cholerae* genome, following a strict set of rules (explained in section 5.2.1.1). These canonical SNPs were checked for their robustness in reconstructing a tree congruent to the original whole genome phylogeny. The SNPs were selected in such a way that the set should be able to withstand the expansion of the phylogeny and be easily customized to answer different questions depending on the resolution required.

The ability to sequence whole bacterial genomes is still limited in many regions of the world where cholera is still occurring. Consequently, methodologies based on simple molecular biology techniques available in routine laboratories that are expandable,

relatively economic and reliable were developed to detect these SNPs. Multiplex ligation-dependent probe amplification (MLPA) (MRC-Holland) can be used to detect specific DNA fragments, insertions, deletions, as well as individual SNPs, permiting the detection of multiple targets by amplification driven by a single primer pair in a convenient single reaction.

The use of MLPA technology requires only a very basic laboratory set up, a thermocycler and gel electrophoresis equipment. MLPA has a simple working mechanism (Figure 5.1), which involves a 5-step reaction that can be performed in a single tube. The first step involves denaturation of sample DNA and hybridisation of MLPA probes. The second step is a ligation reaction, the third is PCR amplification of the probes that have been ligated, the fourth step involves running and separating the amplification products by electrophoresis and the final step is data analysis.

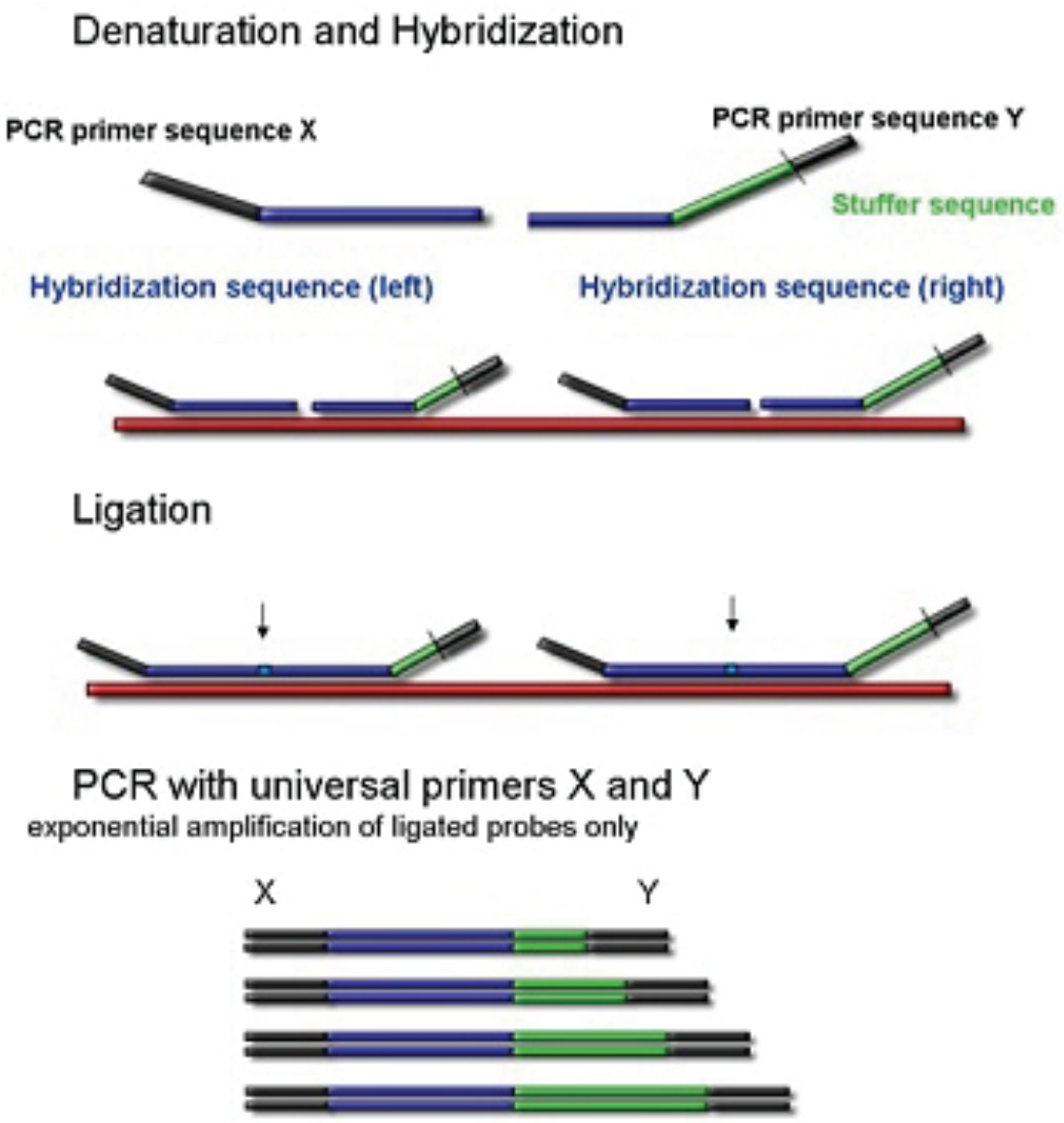


Figure 5.1: Step-by-step guide to MLPA. The targeting hybridization sequence is allele specific and the stuffer sequence is different in length for different probes. PCR using primers for universal X and Y primer sequences will give different length amplicons for different alleles. After the template DNA denaturation, the two probe oligonucleotides are both hybridised to their adjacent targets so that they can be ligated, which is successful only when the desired allele is present. Then, a PCR with universal primers amplifies the successfully ligated probes. The intensity of bands on agarose gel is then used to represent the number of target allele sequences present in the sample. Figure sourced from http://www.mlpa.com/.

Since MLPA does not amplify the target DNA but MLPA probes that hybridise to the target sequence, only a single pair of PCR primers is required per reaction. After DNA denaturation, the two probe oligonucleotides are both hybridised to their adjacent targets so that they can be ligated. However, ligation of both the probes is successful only if the desired allele is present. Thus, subsequent PCR only generates amplified DNA products for the ligated probes. Failure to amplify a product indicates the lack of the targeted SNP. The intensity of bands representative of particular amplification products is indicative of the number of amplified products, corresponding to the number of target allele sequences in the sample. With this technology, the removal of unbound probes is not required, which means that the process is even more streamlined, ideal for the basic molecular biology laboratories.

This chapter describes the design, utility and evaluation of the robustness of a novel SNP based typing scheme for *V. cholerae*. Initially, the rationale behind the selection of SNPs is discussed. Secondly, the impact of the addition of more sequenced *V. cholerae* from the seventh pandemic lineage on to the tree and any impact on SNP selection is detailed. Finally, the design of two MLPA kits with different potential to interrogate the phylogeny of *V. cholerae* is discussed.

5.2 Results and discussion

5.2.1 SNPs for genotyping

5.2.1.1 Selection of canonical SNPs

As the first step towards designing a simplified SNP typing assay for *V. cholerae* El Tor isolates, 27 SNPs were identified, which could be used to reconstruct the seventh pandemic phylogeny (Figure 5.2). The selection was made from the 1757 SNPs identified as described in chapter 2 (Mutreja *et al,* 2011), where DNA from 122 *V. cholerae* El Tor was sequenced and analysed. This original dataset showed that this *V. cholerae* population evolves at a mutation rate of 3.3 SNPs/year. Capitalizing on this unique, slow and clock-like evolutionary rate, the 27 canonical SNPs were carefully selected from well-defined branches of the tree.

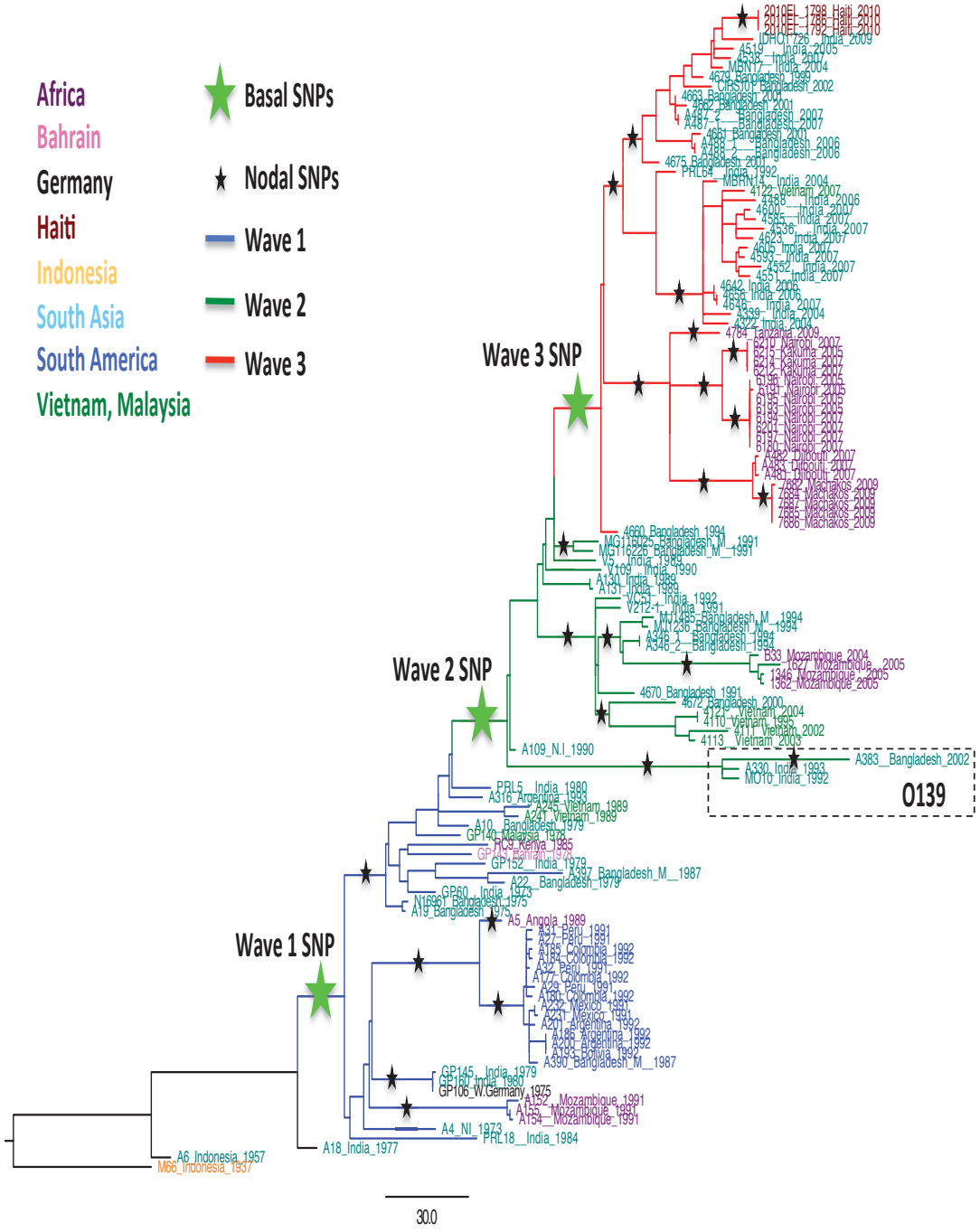


Figure 5.2: A maximum likelihood phylogenetic tree for the global seventh pandemic lineage L2 *V. cholerae* as described in chapter 2. A key to the SNPs selected (represented by stars) for typing is provided in the figure and the scale represents the number of SNPs.

Each of the SNPs was selected after checking against a strict set of rules to make the SNP typing scheme robust and expandable. SNPs that were filtered out did not fit the

following criteria: 1) the SNP should not be from potentially mobile genomic region such as CTX, pathogeneticity islands or the super-integron; 2) non-snonymous were given preference over synonymous SNPs, as there is arguably less chance of reversion to the more common allele and least preference was given to intergenic SNP; 3) the SNP selected should not be homoplasic or under likely selection pressure; 4) the SNP selected should not be an obvious consequence of recombination (homologous or recombinase driven); 5) in the complete dataset of 1757 SNPs, there should be no other SNP within 1kb 3' or 5' of the selected SNP; 6) SNPs on chromosome 1 were preferred over chromosome 2 SNPs as most of the 'house keeping' genes are on chromosome one. For each of the finally selected 27 SNPs, there were 3 to 5 more alternate SNPs identified in the genome, contributing to the same branch and following the exact criterion.

The selected SNPs were divided into basal (bSNPs) or nodal (nSNPs) SNPs, to make the typing assay flexible and capable of answering questions at different levels of resolution. Each of the three selected basal non-synonymous SNPs was specific to one of the three cholera seventh pandemic waves. A reconstruction of the phylogeny of the 122 original *V. cholerae* El Tor described in chapter 2 is shown in Figure 5.3. The wave-1 defining SNP, an A>T transversion, was at base pair position 996160 within VC_0930 encoding for a hemolysin-related protein. The wave-2 differentiating SNP, a T>C transition, was at base pair position 716076 within VC_0668 encoding for DNA mismatch repair protein, MutH. The wave-3 differentiating SNP, an A>G transition, was at base pair position 427292 within VC_400 encoding for MSHA biogenesis protein, MshJ. The SNPs were identified against the *V. cholerae* N16961 reference El Tor genome (Heidelberg, *et al.*, 2000) and they are listed in Table 5.1 in blue.

The 24 high-resolution nSNPs (coloured red in Table 5.1) including 4 synonymous, 1 intergenic and 19 non-synonymous were selected to provide a deeper resolution into the seventh pandemic phylogeny. These SNPs were able to further classify the wave-1, 2 or 3 isolates into important sub-clades such as West African South American clade (WASA), the O139 serogroup, India-Bangladesh wave-1 or 2 or 3, Kenyan, Mozambique wave-1 or 2, Matlab, Vietnam wave-1 or 2, Haiti and others. Questions that could be answered using these 27 SNPs are listed in the Table 5.2.

SNP No.	Syn/Non_Syn/Int	SNP Detail	Position (bp)	Strand
1a	Non-Syn	A->T	996160	Forward
1b	Non-Syn	C->T	1587464	Forward
1c	Non-Syn	C->T	1861198	Forward
1d	Non-Syn	G->A	2667703	Reverse
1e	Non-Syn	C->T	3995581	Forward
2a	Non-Syn	T->C	716076	Reverse
2b	Non-Syn	C->A	775752	Reverse
2c	Non-Syn	A->T	1401879	Reverse
2d	Non-Syn	C->T	2234269	Reverse
2e	Non-Syn	T->C	2295509	Reverse
3a	Non-Syn	A->G	427292	Forward
3b	Non-Syn	G->A	432681	Forward
3c	Non-Syn	G->A	1368686	Reverse
3d	Non-Syn	G->A	1641070	Forward
3e	Non-Syn	A->C	1809138	Reverse
1a	Non-Syn	C->T	68893	Reverse
1b	Non-Syn	A->G	116786	Forward
1c	Non-Syn	C->T	554545	Reverse
1d	Non-Syn	G->A	698661	Reverse
1e	Non-Syn	A->C	758025	Forward
2a	Non-Syn	C->T	1221186	Reverse
2b	Non-Syn	A->G	1933622	Forward
2c	Non-Syn	A->C	2269996	Reverse
2d	Non-Syn	G->A	2450006	Reverse
2e	Non-Syn	C->T	3657508	Reverse
3a	Non-Syn	C->T	34254	Reverse
3b	Non-Syn	C->T	286337	Forward
3c	Non-Syn	C->T	720770	Forward
3d	Non-Syn	A->G	1921705	Reverse
3e	Non-Syn	G->A	1989285	Forward
4a	Non-Syn	C->T	847179	Reverse
4b	Syn	T->A	1889275	Forward
4c	Syn	C->A	2180993	Forward
4d	Non-Syn	G->T	2914537	Reverse
5a	Non-Syn	C->T	27804	Reverse

5b	Non-Syn	C->A	364550	Reverse
5c	Non-Syn	G->A	2352451	Forward
5d	Non-Syn	T->C	3159854	Forward
5e	Non-Syn	G->A	3581904	Forward
6a	Non-Syn	G->A	1116888	Forward
6b	Non-Syn	G->T	1728150	Reverse
6c	Syn	C->T	2044594	Forward
6d	Non-Syn	G->T	3231581	Reverse
6e	Non-Syn	G->A	3446126	Forward
7a	Non-Syn	G->A	103247	Forward
7b	Non-Syn	C->T	918801	Reverse
7c	Non-Syn	G->A	1513879	Forward
7d	Non-Syn	G->A	1838836	Reverse
7e	Non-Syn	C->T	2766132	Reverse
8a	Non-Syn	A->G	822442	Forward
8b	Non-Syn	C->T	1763485	Forward
8c	Non-Syn	A->C	1978660	Reverse
8d	Non-Syn	G->A	2012227	Reverse
8e	Non-Syn	C->T	2088750	Forward
9a	Non-Syn	C->T	748559	Forward
9b	Non-Syn	G->A	3019906	Reverse
9c	Non-Syn	C->T	3448829	Forward
9d	Non-Syn	C->T	3757907	Reverse
9e	Non-Syn	C->T	3980112	Reverse
10a	Non-Syn	A->G	2734994	Forward
10b	Intergenic	C->T	3175627	Reverse
10c	Syn	C->A	3945042	Reverse
11a	Non-Syn	A->T	819598	Reverse
11b	Non-Syn	T->A	1373908	Reverse
11c	Non-Syn	G->A	1775827	Reverse
12a	Non-Syn	A->T	62257	Forward
12b	Non-Syn	G->A	203857	Forward
12c	Non-Syn	C->A	333332	Forward
12d	Non-Syn	C->T	632341	Forward
12e	Non-Syn	C->T	641983	Forward
13a	Non-Syn	G->A	252140	Forward
13b	Non-Syn	T->A	322738	Forward

13c	Non-Syn	T->C	368120	Forward
13d	Non-Syn	C->T	2059765	Forward
13e	Non-Syn	T->C	2619351	Reverse
14a	Non-Syn	A->T	89430	Reverse
14b	Non-Syn	G->A	760185	Forward
14c	Non-Syn	C->T	1395635	Reverse
14d	Non-Syn	C->A	1417110	Reverse
14e	Non-Syn	G->A	2336701	Reverse
15a	Syn	G->A	388326	Forward
15b	Non-Syn	G->T	952983	Forward
15c	Non-Syn	T->A	1097210	Forward
15d	Non-Syn	C->T	2249858	Reverse
15e	Syn	G->A	2290774	Reverse
16a	Non-Syn	G->T	652539	Reverse
16b	Non-Syn	C->T	1833923	Forward
16c	Non-Syn	G->A	2057242	Reverse
16d	Syn	A->C	2309742	Reverse
16e	Non-Syn	T->G	2584695	Reverse
17a	Syn	C->T	426049	Forward
17b	Non-Syn	G->A	1129403	Reverse
17c	Non-Syn	C->T	1472551	Reverse
17d	Non-Syn	A->C	1795218	Forward
17e	Non-Syn	G->A	2124622	Reverse
18a	Non-Syn	C->T	430898	Forward
18b	Syn	C->T	2257626	Reverse
18c	Syn	G->A	2334969	Reverse
18d	Non-Syn	C->T	2742849	Reverse
18e	Intergenic	A->C	3669444	Forward
19a	Non-Syn	C->A	659772	Reverse
19b	Non-Syn	G->T	1241084	Reverse
19c	Non-Syn	C->T	2122620	Reverse
19d	Non-Syn	C->T	2625954	Forward
19e	Non-Syn	T->C	3810829	Forward
20a	Non-Syn	G->T	290017	Reverse
20b	Intergenic	C->T	1043159	Forward
20c	Non-Syn	C->T	2276983	Reverse
20d	Non-Syn	C->A	2968947	Reverse

21a	Intergenic	G->T	710243	Forward
21b	Syn	G->A	1827217	Reverse
21c	Non-Syn	G->A	3104942	Forward
22a	Non-Syn	A->G	991452	Forward
22b	Non-Syn	C->T	1344021	Forward
22c	Syn	C->T	1961296	Forward
22d	Syn	A->G	2242015	Forward
22e	Non-Syn	T->G	3122273	Forward
23a	Syn	G->A	72585	Forward
23b	Non-Syn	G->A	1782519	Reverse
23c	Syn	C->T	2203923	Forward
23d	Intergenic	G->T	2669958	Forward
23e	Syn	C->T	3384140	Forward
24a	Syn	T->C	139591	Forward
24b	Non-Syn	C->T	148860	Reverse
24c	Non-Syn	C->A	2098556	Forward
24d	Non-Syn	C->T	2577815	Reverse
24e	Non-Syn	C->A	3538244	Reverse

Table 5.1: Table showing the list of 27 SNPs selected for genotyping from the original seventh pandemic *V. cholerae* tree. Those coloured blue are bSNPs or basal SNPs and in red are nSNPs or nodal SNPs.

Resolution provided by the 27 SNPs
1) Does the strain belong to wave-1 ?
2) Does the strain belong to wave-2 ?
3) Does the strain belong to wave-3 ?
1) Is the isolate wave-1 Mozambique ?
2) Is the isolate from the West Germany Group?
3) Does the isolate belong to WASA-1 cluster?
4) Is the isolate WASA1/West African ?
5) Does the isolate belong to WASA-1/South American ?
6) Is the isolate just wave-1 but not WASA-1, from the West German cluster or Mozambique ?
7) Is the isolate from the O139 lineage ?
8) Is the isolate of A383 strain type of the O139 lineage ?
9) Is the isolate wave-2 Mozambique, Matlab or Vietnamese ?
10) Is the isolate from wave-2 Vietnam cluster ?

11) Is the isolate from wave-2 Mozambique and Matlab cluster ?
12) Is the isolate wave-2 Mozambique ?
13) Is the isolate wave-2 Matlab outside Mozambique and Matlab cluster ?
14) Is the isolate from wave-3 Kenyan cluster ?
15) Is the isolate from Tanzania lineage ?
16) Is the isolate from Nairobi and Kakuma cluster ?
17) Is the isolate from Nairobi and Kakuma mixed cluster ?
18) Is the isolate from Nairobi cluster ?
19) Is the isolate from Djibouti and Machakos cluster ?
20) Is the isolate from Machakos cluster ?
21) Is the isolate from India, Bangladesh or Haiti ?
22) Is the isolate not from Haiti + India Bangladesh cluster ?
23) Is the isolate from Haiti + India Bangladesh cluster ?
24) Is the isolate from Haiti ?

Table 5.2: Table listing 27 questions that may be addressed by using the 3 bSNPs (in blue) and 24 nSNPs (in red) selected for genotyping.

5.2.1.2 Phylogenetic analysis on selected SNPs

The positions of the 27 informative and canonical SNPs were used to reconstruct the alignment of the whole genomes of the originally sequenced 122 *V. cholerae* described in chapter 2. A maximum likelihood tree was generated using the resolution information of these SNPs. Each of the selected SNP represented a different branch on the tree and was sufficient to differentiate isolates to different braches of the tree.

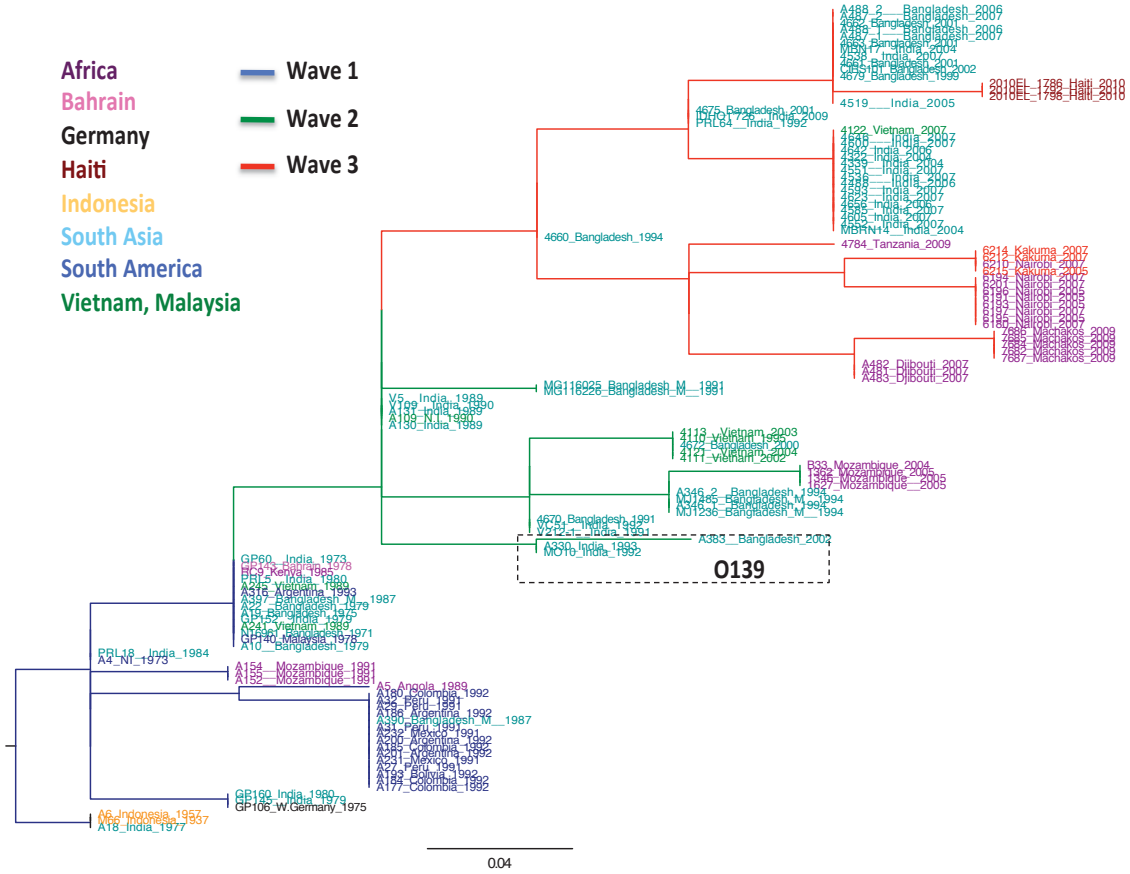


Figure 5.3: A maximum likelihood phylogeny of the 122 *V. cholerae* of L2 lineage described chapter 2, based on the 27 selected SNPs. Each branch is based on a single SNP. A key to locations from which the isolates originated from and the waves to which they belong is provided in the figure and the scale given represents substitutions per variable site.

5.2.2 Phylogeny expansion and MLPA kits

5.2.2.1 Global dissemination of wave-3 in 3 sub-waves

To expand the number of isolates mapped to the phylogenetic framework built as described in chapter 2, DNA from 802 new *V. cholerae*, spanning more than 50 years of dates of isolation, was sequenced. In addition, the publically available data for 200 additional sequences obtained from the NCBI and EBI databases were incorporated into the analysis, totaling 1002 seventh pandemic *V. cholerae*. The accession numbers of all the isolates are provided in Appendix.

All the seventh pandemic isolates sequenced at the WTSI were *V. cholerae* O1 El Tor or O139 originally from patients with clinical disease. A high resolution, maximum likelihood phylogeny based on genome wide SNPs was constructed using the methods previously described. The sequence reads were mapped to N16961, a seventh pandemic *V. cholerae* isolated in Bangladesh in 1975, as reference (Heidelberg, *et al.*, 2000). The pre-seventh pandemic isolate M66 (2) was used to root the tree. Mobile genetic elements and genomic islands that are not present in all seventh pandemic *V. cholerae* were not used to call the SNPs. Although relatively limited in the seventh pandemic *V. cholerae*, any SNP dense clusters or likely homologous recombination regions were removed from the analysis (Croucher, *et al.*, 2011) before using the data to construct a consensus tree based on 5335 SNPs (Figure 5.4).

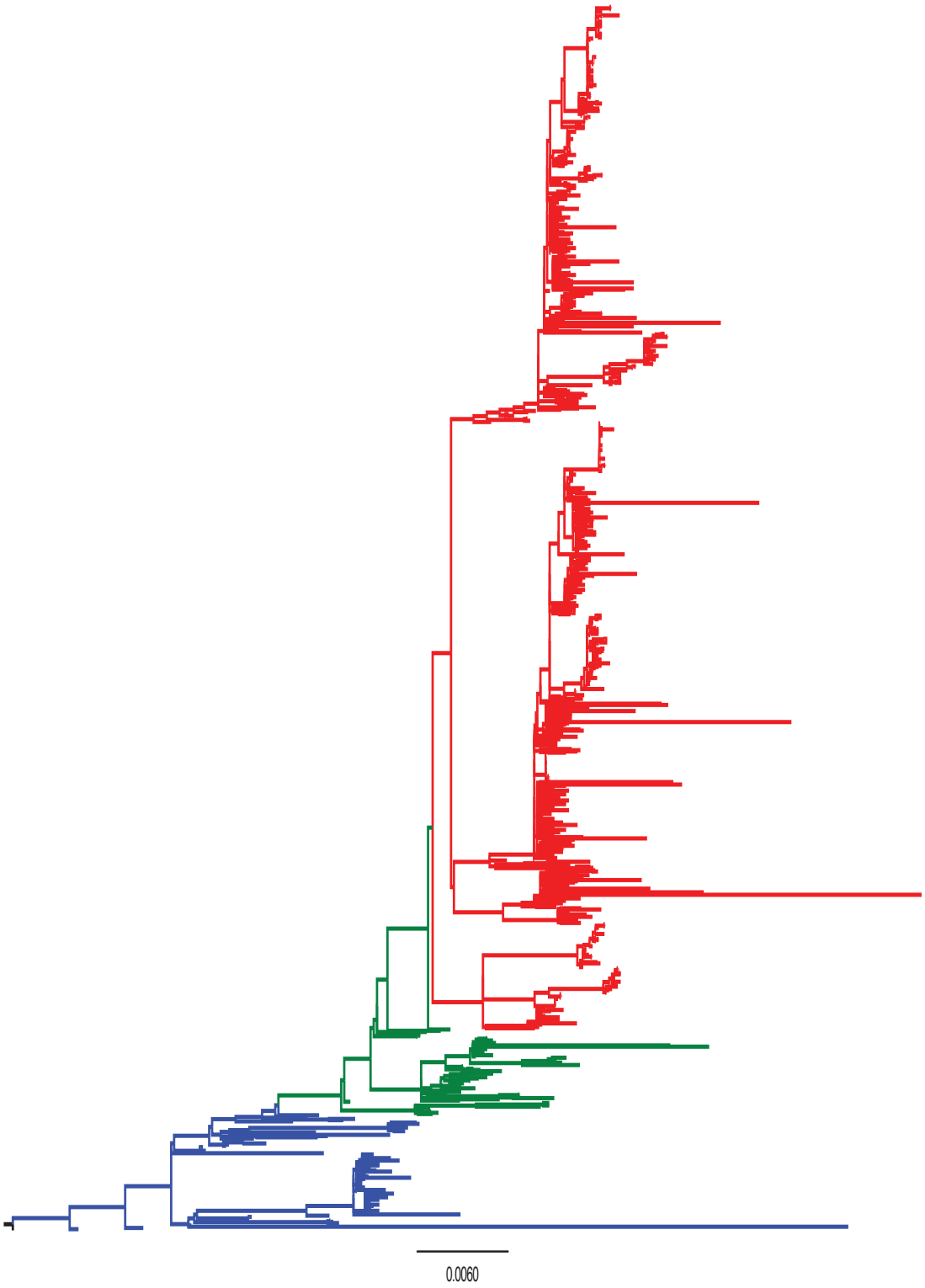


Figure 5.4: The structure of **a** maximum likelihood phylogenetic tree of 1002 seventh pandemic *V. cholerae* El Tor. Blue, green, and red branch colours represent wave-1, 2 and 3 respectively. The reference strain for this tree is *V. cholerae* N16961 El Tor and is rooted using a pre-seventh pandemic *V. cholerae* M66 (2). The scale given is substitutions per variable site.

The shape of the expanded tree based on these 1002 isolates (Figure 5.4) reiterates the monophyletic nature of the seventh pandemic with the older strains at the bottom of the tree or closest to the root and the most recent at the top. Since *V. cholerae* evolves in a strict clock-like manner (Mutreja, *et al.*, 2011), linear regression analysis was performed on this ~8 fold larger collection by plotting the root-to-tip distance of each isolate against its time of isolation. A strong correlation of R^2=0.6 was noted with the rate of substitution per variable site of 0.0006. Interestingly, this provides a nearly identical rate of SNP accumulation rate (3.2 SNPs per year) to that proposed previously (3.3 SNPs per year) (Mutreja, *et al.*, 2011).

In this expanded tree, the new *V. cholerae* were from south-Asia, Africa, Haiti, Gaza, Jerusalem and countries in South America as these were the only places in the world form where cholera has been reported in rbjobsecent years (apart from known travellers to these regions). New South American isolates clustered in both wave-1 and wave-2 (chapter 3), isolates from Gaza and Jerusalem clustered in wave-2, whereas all the other isolates added to the original tree were part of wave-3 (Figure 5.5). Moreover, each isolate possessed the signatures regions (chapter 2) such as WASA-1, SXT, GI-11, GI-15, recombination in O-antigen cluster and modifications in VSP-2 or VPI-1 predicted according to their position on the tree. Since most of the new isolates fell into wave-3, this wave naturally attained the highest resolution (Figure 5.5).

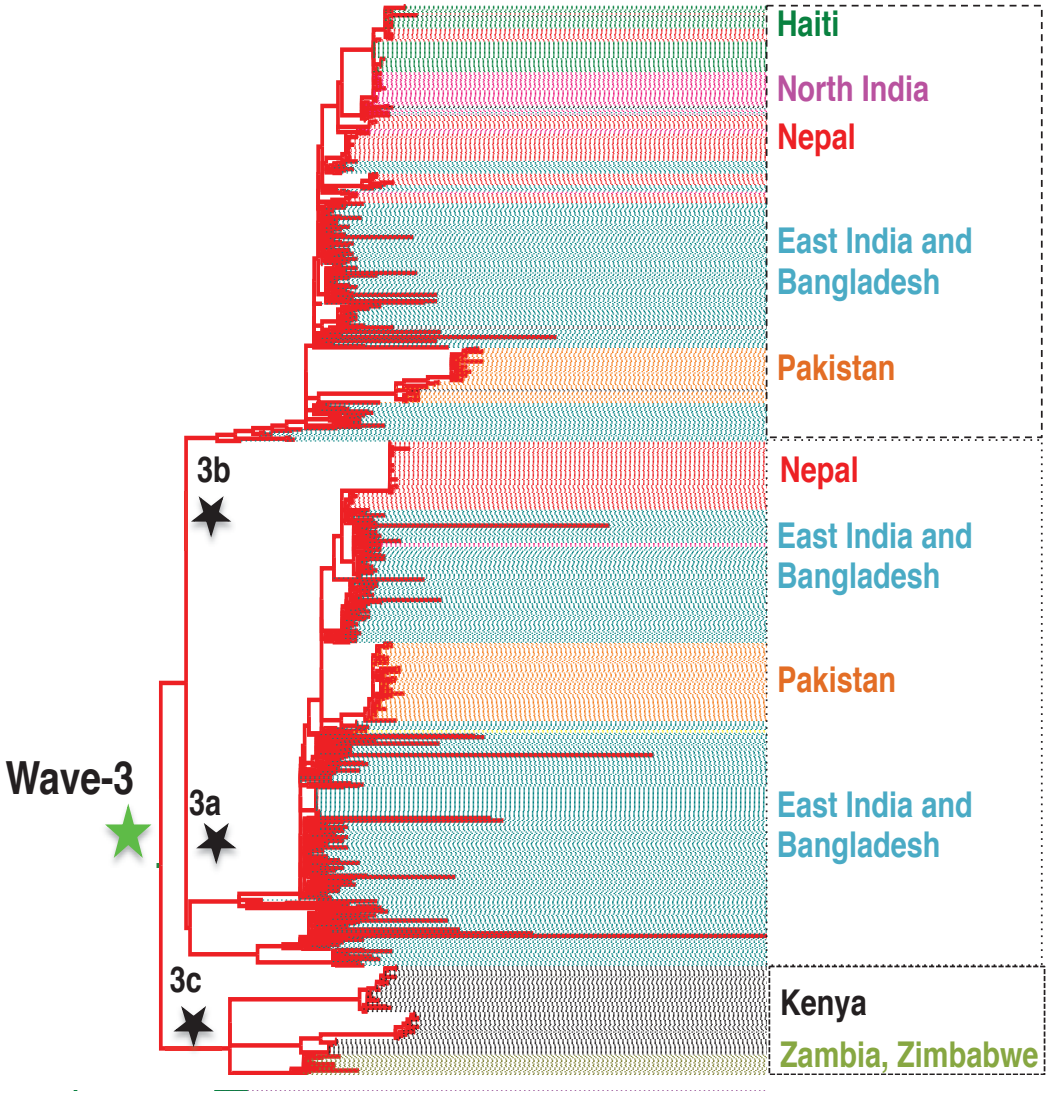


Figure 5.5: A higher resolution image showing more details of the structure of wave-3, which is not obvious in Figure 5.4. The majority of the more recently isolated *V. cholerae* fell in wave-3 as cholera moved through parts of African, Haitian and south Asian, providing a deeper phylogenetic structure. The division of wave-3 into 3a, 3b and 3c is apparent as African isolates fell in a single clade (3c) and south Asian and Haitian isolates fell in two other sub-clades (3a and 3b).

After the analysis based on the additional *V. cholerae* isolates it became clear that wave-3 could be sub-divided into three major sub-clades now referred to as 3a, 3b and 3c (Figure 5.5). The isolates in sub-clades 3a and 3b were primarily from the south-Asian sub-continent whereas isolates from Haiti all fell in sub-clade 3b. Isolates from Kenya (described in more detail in chapter 3), Zambia and Zimbabwe clustered as one

lineage in sub-clade 3c.

5.2.2.2 Design of the MLPA based SNP-genotyping assays

Since the expanded seventh pandemic phylogeny provided a deeper structure and showed regional sub-clades that were previously not obvious, new SNPs were selected to discriminate within these sub-lineages, including wave-3 sub-clades 3a, 3b and 3c, Pakistan sub-clades 1 and 2 and the Haitian sub-clade. Alongside these SNPs, MLPA probes were also designed based on genome differences that were not simple SNPs (Chiang, *et al.*, 2006; Chun, *et al.*, 1999; Garza, *et al.*, 2012; Hoshino, *et al.*, 1998; Ramachandran, *et al.*, 2007). For example, probes were designed for resolving *V. cholerae* from the other commonly found *Vibrio* species *V. mimicus*; for differentiating *V.cholerae* serogroups O1/O139 from others; for detecting islands within the seventh pandemic lineage such as WASA-1 and SXT; for differentiating the traditional *ctx*B gene types. The probes were divided into two kits with the Kit -1 designed for routine *V. cholerae* typing and Kit-2 for users who are interested in higher resolution once they have established through Kit -1 or other means that the sample they are working on was *V. cholerae* O1 El Tor. The compositions of both kits are given in Table 5.3 below.

MLPA Kit - 1		
Probe	**Positive for/Resolution**	**L**
p1t	Wave-1	130
p4c	ctxB classical	154
O139_fw/rv_primers	O139	178
s14t	W3c (Kenya)	202
O1_fw/rv_primers	O1	226
p6a	ctxB-3b	250
WASA_fw/rv_primers	WASA-1 island	282
SXT_fw/rv_primers	SXT	314
p3g	Wave-3	346
s3t	Wave-1 WASA cluster	378
Vch_fw/rv_primers	*V. cholerae/V. mimicus*	418
p2c	Wave-2	458
Vsp_fw/rv_primers	*Vibrio* sp. (positive control)	498

MLPA Kit - 2		
Probe	**Positive for/Resolution**	**L**
p1t	Wave-1 (positive control)	130
s22g	Wave-3a (South Asian clade)	154
s23a	Wave-3b (South Asian + Haitian clade)	178
s14t	W3c (Kenya)	202
s25t	Pakistan SC-1	226
p6a	ctxB-3b	250
WASA_fw/rv_primers	WASA-1 island	282
s26c	Pakistan SC-2	314
s27a	Nepal-Haiti clade	346
s3t	Wave-1 WASA cluster	378
s5t	Latin American in Wave - 1	418
p2c	Wave-2	458
Vsp_fw/rv_primers	*Vibrio* sp. (positive control)	498

Table 5.3: Table showing the primer design of the MLPA kits. Column 1 shows whether the probe was designed using SNP or primer information. The regions these probes identify are listed in column 2 and the size of the product expected for each probe amplicon is given in column 3.

The kits are currently under construction at MLPA Holland and will be validated on *V. cholerae* test samples that have already been sent to their laboratory. These samples are anonymous to them but known to us.

5.3 Lessons learned from the expanded phylogeny and importance of SNP genotyping

The addition of new *V. cholerae* isolates to the seventh pandemic phylogeny consolidated the view that this is a highly monophyletic lineage. The linear regression analysis again showcased the clock-like evolution of the seventh pandemic *V. cholerae* O1 El Tor. When the new phylogenetic data was plotted onto the world map (Figure 5.6), the previously postulated pattern of spread of the seventh pandemic was conserved. All new clades and branches of the expanded phylgeny also radiated from the same source population, strengthening the previously proposed hypothesis that

there is a single backbone population that is seeding cholera globally in the form of independent but overlapping waves that enter an area, cause cholera outbreaks that then die out when a new wave arrives. However, wave-1 *V. cholerae* that have apparently persisted in Mexico (chapter 3) are an exception to this hypothesis, and evidence for this persistence of this clade should be investigated across the regions of Latin American affected by the 1990s cholera outbreaks.

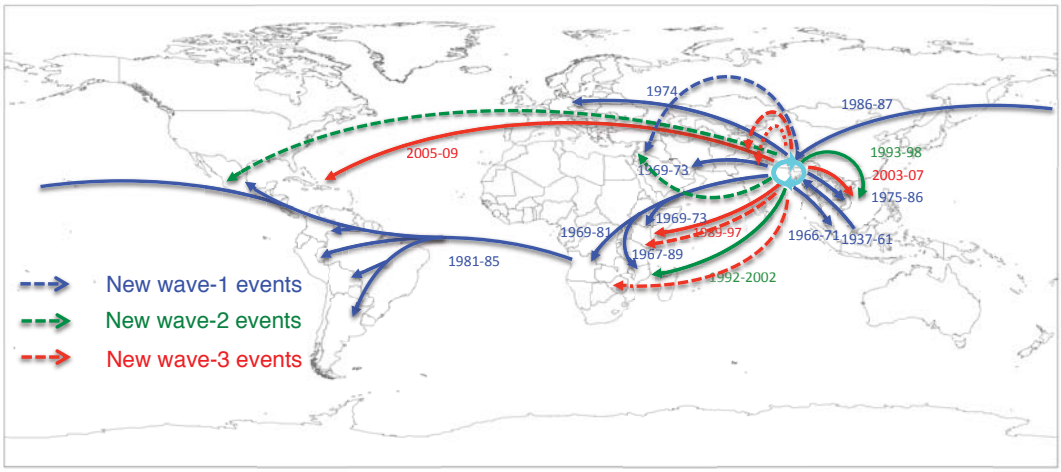


Figure 5.6: Figure showing the proposed spread of newly added *V. cholerae* isolates in the expanded seventh phylogenetic study. The dashed lines represent the newly plotted events and solid lines represent previously published data (Mutreja, *et al.*, 2011).

Examining the structure of wave-3 in finer detail, it is clear that only sub-clade 3b isolates entered and spread to Haiti from south-Asia. Since no sub-clade 3a isolates have been reported in Haiti even after 2 years of thorough surveillance, this suggests that the *V. cholerae* seeding source in Haiti was from within a relatively confined regional boundary. To trace the source to an exact location, a thorough GPS coordinated study of *V. cholerae*, collected just before the Haitian cholera outbreak from the south Asian countries including the regions close to Nepal, would be needed.

The traditional typing techniques used to define *V. cholerae* are not suitable for investigating the spread and epidemiology of cholera outbreaks and analysis based on whole genome sequences requires considerable resources and informatics skill. This is where the MLPA Kit based SNP genotyping approach described here could fill the gap and could provide robust answers based on the growing *V. cholerae* sequence

databases. By using these kits, quick and reliable information about any isolates' position in the global seventh pandemic phylogenetic framework could be gained. This information could provide vital clues to facilitate tackling the spread of cholera. Even if an isolate cannot be mapped to the phylogeny using the proposed SNP genotyping, such a result could signify the emergence or identification of a new phylotype. In such an event, other public health centers would be alerted in advance, so their monitoring programs could look for such new clades and efforts to contain their spread could be initiated.

6. Conclusion and Future Directions

6.1 Conclusion

Even though cholera is regarded in many quarters as an old disease, the current global burden is between 3-5 million cases and more than 100,000 deaths are reported worldwide annually (*http://www.who.int/wer*). While many of these cholera cases are in the endemic regions of the Indian subcontinent, the increased incidence of cholera globally since 2007 has highlighted the need of more public health research on this potentially easy to treat disease. The etiological agent, *V. cholerae* is a genetically diverse species but out of more than 200 O-antigen serogroups only isolates of O1 and O139 can cause epidemic cholera. Of the two biotypes 'El Tor' and 'classical' of serogroup O1 *V. cholerae*, El Tor account for almost all of the cases in the current seventh pandemic.

This thesis utilizes the whole genome sequencing data generated from more than 1000 seventh pandemic *V. cholerae* El Tor as well as isolates from non-seventh pandemic lineages to prove that the classical and El Tor biotypes evolved independently and do not share a recent common ancestor. This data also clearly showed that currently successful lineage of El Tor isolates is monoclonal or monophyletic and evolves in a strict clock-like manner. The phylogeographical analysis indicates that in the seventh pandemic cholera has spread in the form of independent but overlapping waves from a source population in the Bay of Bengal to other regions of the world (chapter 2).

The total *V. cholerae* collection used in this study included sets of smaller collections obtained from different cities, regions or countries on which focused studies could be performed. In chapter 3, a series of case studies described evolutionary patterns identified within populations of *V. cholerae* O1 El Tor obtained from particular countries. Studies on *V. cholerae* isolated during the catastrophic 2010 floods in Pakistan identified two clear introductions of the disease into the country and provided insight into their spread within Pakistan. The study on *V. cholerae* collected over several decades in Mexico provided evidence for the persistence of wave-1 El Tor isolates since the 1990s. A Kenyan surveillance study shed light on the local

clonality of isolates and provided evidence for a recent common ancestor with South Asian isolates. Again, evidence was provided for independent entries of cholera into Kenya and the presence of two sub-clades.

The large amount of data generated from sequencing over 1000 *V. cholerae* was exploited to design SNP typing assays, in the form of kits that are suitable for use in public health and scientific laboratories in developing countries (chapter 5). Robust SNPs were selected, based on phylogenetic analysis, and the kits were designed so that they could rapidly and accurately identify *V. cholerae* associated with any outbreak.

The genotypic basis of the Ogawa to Inaba serotype conversion in *V. cholerae* O1 El Tor was studied in detail by analysing the *wbe*T sequence from 777 of the 1002 seventh pandemic *V. cholerae* analysed in this study. This analysis identified the mutations that underpinned the serotype switching and provided insight into mechanism. *V. cholerae* collected during a phase III vaccine trial in Kolkata, India were examined to identify any temporal and serotypic correlation in a phylogenetic context (chapter 4). This analysis provided evidence for regular sweeps of Inaba and Ogawa types spreading through the trial sites over successive cholera seasons. Such information is likely to have value for supporting future vaccine studies in the field.

In summary, the data described in this PhD thesis may facilitate future cholera surveillance performed as part of public health programs at a local or national level, facilitating quick and directed actions to contain the spread of an outbreak. The academic community will also benefit from this data, which is publically available to further research on the biology and epidemiology of *V. cholerae*.

6.2 Future Directions

6.2.1 Further expansion of the sequenced *V. cholerae* collection

Perhaps the most obvious way to extend this work is to expand efforts on the whole genome based phylogenetic analysis of the seventh pandemic collection by the

addition of new *V. cholerae* isolates from around the globe. Through the existing and new collaborations between the WTSI and partners based around the world, such an effort is already well in progress. Recent and historical strain collections from Pakistan, West Africa, India and Bangladesh are being sequenced to add finesse and detail to the structure of the global seventh pandemic phylogeny. With the addition of new isolates collected with detailed metadata, a more accurate prediction of the spread of *V. cholerae* could be made alongside an increased understanding of the overall phylogenetic framework. For example, a collection of O139 *V. cholerae* has been identified at the National Institute of Cholera and Enteric Diseases (NICED) in Kolkata and arrangements are being made to have these sequenced. The O139 *V. cholerae* appear to fall into a specific lineage within the El Tor phylogenetic tree and such isolates were predominantly isolated between 1992 and 2005, mainly in Kolkata and Bangladesh. The possible reasons behind the success of a lineage for just a few years before it disappeared completely are presently unknown; this project may shed light onto some of the contributing factors. For example, do all O139 isolates fall into the same lineage and do any isolates show evidence of more extensive recombination beyond the known O-antigen loci? This study should highlight the SNP variation that occurred in O139 isolates from the time when they were first causing outbreaks to the time they went extinct. This data may give some insight into any negative selection pressure that these strains may have encountered.

Further, analysis of *V. cholerae* that fall outside the seventh pandemic El Tor lineage could shed more light onto the *V. cholerae* pool circulating in the environment that is clearly interacting with the epidemic *V. cholerae* populations. Detailed analysis of new lineages identified in this PhD (for example, MLE-1 and MLE-2 in section 3.3.8) and previously known lineages such as the US-Gulf coast (section 2.3.1) could provide vital clues about the evolution of vibrios circulating in the environment.

6.2.2 Studies investigating the evolution of *V. cholerae* within cities, countries and continents

While the continued addition of sequenced *V. cholerae* isolates is expanding our understanding of the seventh pandemic, this data is also providing an opportunity to study regional level populations in greater detail. Some of the examples are:

- Post flooding *V. cholerae* continue to be collected in Pakistan and are being analysed to determine further patterns of evolution as well as assess if they are evolving at the same rate as vibrios in other branches of the the seventh pandemic tree;
- New isolates from across Africa are being added to the phylogeny to obtain a representative sample for a pan-continental study and to determine if the dynamic of cholera in Africa is similar to that observed in the representative country, Kenya;
- Studies on cholera in Latin America are being expanded with the inclusion of isolates from Argentina and Brazil. An aim here will be to determine if the persistence of waves in Mexico is common across the whole of Latin America or is just a local phenomenon.

6.2.3 A combined transcriptomics and proteomics study of intestinal tissues taken from mice at different stages of *V. cholerae* infection

In collaboration with researchers at the University of Gothenberg in Sweden, work is currently being conducted to investigate transcriptomics and proteomics patterns in the intestinal tissue of mice during a *V. cholerae* infection. An O1 serogroup *V. cholerae* El Tor Ogawa strain X25049 was used to infect infant mice orally at a dose of 10^6 viable bacteria. Infections were performed and tissues were collected in Sweden from groups of mice at 4 and 18 hours after infection, with uninfected mice acting as controls. RNA-seq and proteomic analysis is currently underway at the WTSI using these materials. The patterns of gene expression will be compared between the infected and un-infected mice and between the mice at different time points. Protein extracts will be analysed using a Liquid Chromatography-Mass Spectrometry (LC-MS) platform to compare the translation differences at the same time points as the RNA analysis. The data from these experiments would be crucial in highlighting aspects of the host response to *V. cholerae* infection. A comparison between transcriptomics and proteomics analyses results should provide a list of genes that could be important candidates in future cholera vaccine or drug design. A further study is also planned where similar mice will be challenged orally with

cholera toxin in order to compare the impact of *V. cholerae* infection with exposure to the cholera toxin.

6.2.4 A study designed to investigate household and community level spread of *V. cholerae*

A study has been instigated to exploit the high sensitivity and resolving power of SNP based phylogenetic analysis of *V. cholerae* genomes, to investigate transmission patterns involving household contacts and index cholera patients. To achieve this, a unique *V. cholerae* collection from Dhaka, Bangladesh will be sequenced at the whole genome level in an attempt to establish transmission chains, which could not be established using VNTR in a previously published study (Kendall, *et al.*, 2010). The *V. cholerae* isolates in this study were collected as follows: In Dhaka, if an individual reported to a hospital associated with the International Centre for Diarrheal Disease and Research (ICDDRB) and was subsequently found to be positive for *V. cholerae*, the patient was classed as the index patient and their household was investigated in order to recruit individuals who shared the same food and water sources. Daily rectal swabs were then taken for 10 days from implicated co-inhabitants. These swabs were then analysed for *V. cholerae* and culture positive samples were stored. With ~100 samples from the index patients and ~150 from the household contacts, this study will use the power of whole genome analysis and detailed metadata in an attempt to identify evidence for transmission within a house or between houses in a community.

In conclusion, the data and analysis provided and described in this thesis is underpinning a series of on going investigations into the biology of *V. cholerae* in both experimental laboratory and field settings.

Methods

Genome sequencing

Genomic DNA for all the *V. cholerae* analysed in this study were extracted by our collaborators and shipped to the Wellcome Trust Sanger Institute (WTSI) for whole genome sequencing. Multiplex sequencing libraries of 250 bp insertion size were created for each sample using the manufacturer's protocol by the sequencing team at WTSI. The libraries were loaded on to the Illumina's GA II or HiSeq platform cell to perform 54-72-base paired-end sequencing of 12-96 separate libraries in each lane. Each library had a unique index tag and after sequencing this tag sequence information was used for assigning reads to the individual samples assisting the downstream separation of the data for each sample. All the samples achieved an average coverage of 50-200x in the regions where SNPs were called. All the data has been submitted to European Nucleotide Archive and the accession codes are listed in the strain tables provided in the chapters.

Whole genome alignment and detection of SNPs

The paired-end read data obtained was mapped to the O1 El Tor reference N16961 (for chromosome 1 and 2, the NCBI accession numbers are AE003852 and AE003853 respectively) using SMALT (http://www.sanger.ac.uk/resources/software/smalt) to obtain a whole genome alignment for all the strains in this study. For SNP calling, the default settings of nucmer program in the MUMmer package (Kurtz, *et al.*, 2004) were used. No SNPs were called from the reads that either did not map to N16961 or from the regions that were absent from the N16961 reference genome. Strict filtering of the SNPs was performed and any SNP with a quality score less than 30 was excluded. Also, a SNP was considered true only if it was present in at least 75% of the reads at any heterogeneously mapped ambiguous sites. High-density SNP clusters and the possible recombination sites were excluded using the methodology of Croucher *et al.* (Croucher, *et al.*, 2011).

Phylogenetic Analysis

Default settings of RAxML v0.7.4 (Stamatakis, 2006) were used to estimate the phylogenetic trees based on all the SNPs recorded against the reference genome as explained above. The number of SNPs on each branch were calculated by reconstructing all the polymorphic events on the tree using PAML (Yang, 2007). M66 (accession numbers CP001233 and CP001234), a pre-seventh pandemic strain and a well-known out-group for the seventh pandemic strains, was used to root the final seventh pandemic phylogenetic tree (Mutreja, *et al.*, 2011) while other trees were left un-rooted or were midpoint rooted. For visualization and ordering of the nodes, phylogenetic tree reading software Figtree (http://tree.bio.ed.ac.uk/software/figtree/) was used.

Comparative Genomics

A multi-contig draft genome was generated for each sample by assembling the paired end reads using a *de-novo* genome assembly program Velvet v0.7.03 (Zerbino and Birney, 2008). The parameters were set to give the best kmer size and at least 20x kmer coverage. Contigs were ordered using Abacas as per the reference N16961 El Tor complete genome sequence (Assefa, *et al.*, 2009; Heidelberg, *et al.*, 2000). Annotation was transferred from the reference sequence to each ordered draft assembly. Artemis Comparison Tool was used for manual comparison of the assembled genomes (Carver, *et al.*, 2008).

Linear Regression Analysis

The final phylogenetic tree was opened using Path-O-Gen v1.3 (http://tree.bio.ed.ac.uk/software/pathogen) and the root-to-tip distance data for each strain was exported to excel. This data was used to plot a linear regression curve against the year of isolation of the strain. The R-squared correlation, slope and p-values were determined using the inbuilt regression package of R-statistical environment.

Bayesian Analysis

The waves of the seventh pandemic were confirmed using BAPS (Corander, *et al.*, 2008; Corander, *et al.*, 2003). The BAPS analysis was performed on the final SNP alignment obtained after removing the recombination, which contained the unique SNP patterns from the seventh pandemic isolates. The program was run using BAPS individual mixture model and three iterations were performed independently to obtain the most optimal partitioning of the sample.

The tree was reconstructed and the ancestral or nodal dates for the strains were inferred using the Bayesian Markov Chain Monte Carlo framework (Drummond and Rambaut, 2007). The final SNP alignment without recombinant sites was used as the input dataset for BEAST in seventh pandemic each dataset (Drummond and Rambaut, 2007) and the rates of evolution on the branches of the tree were estimated using a relaxed molecular clock (Drummond, *et al.*, 2006), providing the flexibility for the rates of evolution to change amongst the branches of the tree. A coalescent constant population size and a GTR model with gamma correction were used. The results produced from three independent chains of 100 million steps each were sampled every 10,000 steps to maintain homogeneity. The first 10 million steps of each chain were binned. The results of the three chains were combined using Log Combiner, and the maximum clade credibility tree was generated using Tree Annotator software in the BEAST package (http://tree.bio.ed.ac.uk/software/beast/). ESS cut off value of 200 was used for each parameter and convergence was visually confirmed using Tracer 1.5 (http://tree.bio.ed.ac.uk/software/tracer).

References

(1993) Large epidemic of cholera-like disease in Bangladesh caused by *Vibrio cholerae* O139 synonym Bengal. Cholera Working Group, International Centre for Diarrhoeal Diseases Research, Bangladesh, *Lancet*, **342**, 387-390.

(2004) Finishing the euchromatic sequence of the human genome, *Nature*, **431**, 931-945.

(2010) Update: cholera outbreak --- Haiti, 2010, *MMWR Morb Mortal Wkly Rep*, **59**, 1473-1479.

Adler, C.J.*, et al.* (2013) Sequencing ancient calcified dental plaque shows changes in oral microbiota with dietary shifts of the Neolithic and Industrial revolutions, *Nature Genetics*, **45**, 450-455, 455e451.

Ahmed, A.M., Shinoda, S. and Shimamoto, T. (2005) A variant type of *Vibrio cholerae* SXT element in a multidrug-resistant strain of Vibrio fluvialis, *FEMS Microbiology Letters*, **242**, 241-247.

Ahmed, K. and Shakoori, A.R. (2002) *Vibrio cholerae* El Tor, Ogawa O1, as the main aetiological agent of two major outbreaks of gastroenteritis in northern Pakistan, *Journal of Health, Population, and Nutrition*, **20**, 96-97.

Albrecht, M.*, et al.* (2010) Deep sequencing-based discovery of the Chlamydia trachomatis transcriptome, *Nucleic Acids Research*, **38**, 868-877.

Ansaruzzaman, M.*, et al.* (2007) Genetic diversity of El Tor strains of *Vibrio cholerae* O1 with hybrid traits isolated from Bangladesh and Mozambique, *International Journal of Medical Microbiology : IJMM*, **297**, 443-449.

Assefa, S.*, et al.* (2009) ABACAS: algorithm-based automatic contiguation of assembled sequences, *Bioinformatics*, **25**, 1968-1969.

Aury, J.M.*, et al.* (2008) High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies, *BMC Genomics*, **9**, 603.

Baranwal, S.*, et al.* (2002) Role of active efflux in association with target gene mutations in fluoroquinolone resistance in clinical isolates of *Vibrio cholerae*, *Antimicrobial Agents and Chemotherapy*, **46**, 2676-2678.

Barquist, L., Boinett, C.J. and Cain, A.K. (2013) Approaches to querying bacterial genomes with transposon-insertion sequencing, *RNA Biology*, **10**.

Barzilay, E.J.*, et al.* (2013) Cholera surveillance during the Haiti epidemic--the first 2 years, *New England Journal of Medicine*, **368**, 599-609.

Basu, A.*, et al.* (2000) *Vibrio cholerae* O139 in Calcutta, 1992-1998: incidence, antibiograms, and genotypes, *Emerging Infectious Diseases*, **6**, 139-147.

Borroto, R.J. and Martinez-Piedra, R. (2000) Geographical patterns of cholera in Mexico, 1991-1996, *International Journal of Epidemiology*, **29**, 764-772.

Bryant, J.M.*, et al.* (2013) Inferring patient to patient transmission of Mycobacterium tuberculosis from whole genome sequencing data, *BMC Infectious Diseases*, **13**, 110.

Butler, D. (2010) Cholera tightens grip on Haiti, *Nature*, **468**, 483-484.

Butler, S.M. and Camilli, A. (2005) Going against the grain: chemotaxis and infection in *Vibrio cholerae*, *Nature Reviews Microbiology*, **3**, 611-620.

Butler, S.M.*, et al.* (2006) Cholera stool bacteria repress chemotaxis to increase infectivity, *Molecular Microbiol*, **60**, 417-426.

Calain, P.*, et al.* (2004) Can oral cholera vaccination play a role in controlling a cholera outbreak?, *Vaccine*, **22**, 2444-2451.

Carver, T.*, et al.* (2008) Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database, *Bioinformatics*, **24**, 2672-2676.

Chain, P.S.*, et al.* (2009) Genomics. Genome project standards in a new era of sequencing, *Science*, **326**, 236-237.

Chatterjee, S.N. and Chaudhuri, K. (2003) Lipopolysaccharides of *Vibrio cholerae*. I. Physical and chemical characterization, *Biochimica et Biophysica Acta*, **1639**, 65-79.

Chatterjee, S.N. and Chaudhuri, K. (2004) Lipopolysaccharides of *Vibrio cholerae* II. Genetics of biosynthesis, *Biochimica et Biophysica Acta*, **1690**, 93-109.

Chatterjee, S.N. and Chaudhuri, K. (2006) Lipopolysaccharides of *Vibrio cholerae*: III. Biological functions, *Biochimica et Biophysica Acta*, **1762**, 1-16.

Chiang, Y.C.*, et al.* (2006) Identification of Bacillus spp., Escherichia coli, Salmonella spp., Staphylococcus spp. and Vibrio spp. with 16S ribosomal DNA-based oligonucleotide array hybridization, *International Journal of Food Microbiology*, **107**, 131-137.

Chin, C.S.*, et al.* (2011) The origin of the Haitian cholera outbreak strain, *New England Journal of Medicine*, **364**, 33-42.

Chun, J.*, et al.* (2009) 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*, **106**, 15442-15447.

Chun, J., Huq, A. and Colwell, R.R. (1999) Analysis of 16S-23S rRNA intergenic spacer regions of *Vibrio cholerae* and Vibrio mimicus, *Applied and Environmental Microbiology*, **65**, 2202-2208.

Clemens, J.*, et al.* (2011) New-generation vaccines against cholera, *Nature Reviews Gastroenterology & Hepatology*, **8**, 701-710.

Clemens, J.D.*, et al.* (1990) Field trial of oral cholera vaccines in Bangladesh: results from three-year follow-up, *Lancet*, **335**, 270-273.

Colwell, R.R. (1996) Global climate and infectious disease: the cholera paradigm, *Science*, **274**, 2025-2031.

Colwell, R.R. (2000) Viable but nonculturable bacteria: a survival strategy, *Journal of infection and Chemotherapy : Official Journal of the Japan Society of Chemotherapy*, **6**, 121-125.

Corander, J.*, et al.* (2008) Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations, *BMC Bioinformatics*, **9**, 539.

Corander, J., Waldmann, P. and Sillanpaa, M.J. (2003) Bayesian analysis of genetic differentiation between populations, *Genetics*, **163**, 367-374.

Croucher, N.J.*, et al.* (2011) Rapid pneumococcal evolution in response to clinical interventions, *Science*, **331**, 430-434.

Cvjetanovic, B. and Barua, D. (1972) The seventh pandemic of cholera, *Nature*, **239**, 137-138.

Davies, B.W.*, et al.* (2012) Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for V. cholerae virulence, *Cell*, **149**, 358-370.

Davis, B.M. and Waldor, M.K. (2003) Filamentous phages linked to virulence of *Vibrio cholerae*, *Current Opinions Microbiology*, **6**, 35-42.

De, S.N. (1959) Enterotoxicity of bacteria-free culture-filtrate of *Vibrio cholerae*, *Nature*, **183**, 1533-1534.

Dohm, J.C.*, et al.* (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing, *Nucleic Acids Research*, **36**, e105.

Drummond, A.J.*, et al.* (2006) Relaxed phylogenetics and dating with confidence, *PLoS Biology*, **4**, e88.

Drummond, A.J. and Rambaut, A. (2007) BEAST: Bayesian evolutionary analysis by sampling trees, *BMC Evolution Biology*, **7**, 214.

Enzensberger, R.*, et al.* (2005) Mixed diarrhoeal infection caused by *Vibrio cholerae* and several other enteric pathogens in a 4-year-old child returning to Germany from Pakistan, *Scandinavian journal of infectious diseases*, **37**, 73-75.

Faruque, S.M.*, et al.* (2003) Reemergence of epidemic *Vibrio cholerae* O139, Bangladesh, *Emerging Infectious Diseases*, **9**, 1116-1122.

Faruque, S.M., *et al.* (2005) Self-limiting nature of seasonal cholera epidemics: Role of host-mediated amplification of phage, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 6119-6124.

Faruque, S.M. and Mekalanos, J.J. (2003) Pathogenicity islands and phages in *Vibrio cholerae* evolution, *Trends Microbiol*, **11**, 505-510.

Faruque, S.M., *et al.* (2005) Seasonal epidemics of cholera inversely correlate with the prevalence of environmental cholera phages, *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 1702-1707.

Faruque, S.M., *et al.* (2003) Emergence and evolution of *Vibrio cholerae* O139, *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1304-1309.

Faruque, S.M., *et al.* (2000) The O139 serogroup of *Vibrio cholerae* comprises diverse clones of epidemic and nonepidemic strains derived from multiple V. cholerae O1 or non-O1 progenitors, *Journal of Infectious Diseases*, **182**, 1161-1168.

Fasano, A., *et al.* (1991) *Vibrio cholerae* produces a second enterotoxin, which affects intestinal tight junctions, *Proceedings of the National Academy of Sciences of the United States of America*, **88**, 5242-5246.

Finkelstein, R.A. and LoSpalluto, J.J. (1969) Pathogenesis of experimental cholera. Preparation and isolation of choleragen and choleragenoid, *The Journal of Experimental Medicine*, **130**, 185-202.

Flach, C.F., *et al.* (2007) Broad up-regulation of innate defense factors during acute cholera, *Infection and Immunity*, **75**, 2343-2350.

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

Franco, A.A., *et al.* (1997) Cholera in Lima, Peru, correlates with prior isolation of *Vibrio cholerae* from the environment, *American Journal of Epidemiology*, **146**, 1067-1075.

Garza, D.R., *et al.* (2012) Genome-wide study of the defective sucrose fermenter strain of *Vibrio cholerae* from the Latin American cholera epidemic, *PLoS One*, **7**, e37283.

Glass, R.I., *et al.* (1982) Endemic cholera in rural Bangladesh, 1966-1980, *American Journal of Epidemiology*, **116**, 959-970.

Glass, R.I., *et al.* (1985) Predisposition for cholera of individuals with O blood group. Possible evolutionary significance, *American Journal of Epidemiology*, **121**, 791-796.

Griffith, D.C., Kelly-Hope, L.A. and Miller, M.A. (2006) Review of reported cholera outbreaks worldwide, 1995-2005, *The American Journal of Tropical Medicine and Hygiene*, **75**, 973-977.

Harismendy, O.*, et al.* (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies, *Genome Biology*, **10**, R32.

Harris, J.B.*, et al.* (2008) Susceptibility to *Vibrio cholerae* infection in a cohort of household contacts of patients with cholera in Bangladesh, *PLoS Neglected Tropical Diseases*, **2**, e221.

Harris, S.R.*, et al.* (2012) Whole-genome analysis of diverse Chlamydia trachomatis strains identifies phylogenetic relationships masked by current clinical typing, *Nature Genetics*, **44**, 413-419, S411.

Harris, S.R.*, et al.* (2010) Evolution of MRSA during hospital transmission and intercontinental spread, *Science*, **327**, 469-474.

Hasan, N.A.*, et al.* (2012) Genomic diversity of 2010 Haitian cholera outbreak strains, *Proceedings of the National Academy of Sciences of the United States of America*, **109**, E2010-2017.

He, M.*, et al.* (2010) Evolutionary dynamics of Clostridium difficile over short and long time scales, *Proceedings of the National Academy of Sciences of the United States of America*, **107**, 7527-7532.

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

Hendriksen, R.S.*, et al.* (2011) Population genetics of *Vibrio cholerae* from Nepal in 2010: evidence on the origin of the Haitian outbreak, *mBio*, **2**, e00157-00111.

Hernandez, D.*, et al.* (2008) De novo bacterial genome sequencing: millions of very short reads assembled on a desktop computer, *Genome Research*, **18**, 802-809.

Hochhut, B.*, et al.* (2001) Formation of chromosomal tandem arrays of the SXT element and R391, two conjugative chromosomally integrating elements that share an attachment site, *Journal of Bacteriology*, **183**, 1124-1132.

Hochhut, B. and Waldor, M.K. (1999) Site-specific integration of the conjugal *Vibrio cholerae* SXT element into *prfC*, *Molecular Microbiology*, **32**, 99-110.

Holden, M.T.*, et al.* (2013) A genomic portrait of the emergence, evolution, and global spread of a methicillin-resistant Staphylococcus aureus pandemic, *Genome Research*, **23**, 653-664.

Holmgren, J.*, et al.* (1975) Interaction of cholera toxin and membrane GM1 ganglioside of small intestine, *Proceedings of the National Academy of Sciences of the United States of America*, **72**, 2520-2524.

Holmgren, J., Lonnroth, I. and Svennerholm, L. (1973) Fixation and inactivation of cholera toxin by GM1 ganglioside, *Scandinavian Journal of Infectious Diseases*, **5**, 77-78.

Holmgren, J.*, et al.* (1977) Development of improved cholera vaccine based on subunit toxoid, *Nature*, **269**, 602-604.

Holt, K.E.*, et al.* (2012) Shigella sonnei genome sequencing and phylogenetic analysis indicate recent global dissemination from Europe, *Nature Genetics*, **44**, 1056-1059.

Hornick, R.B.*, et al.* (1971) The Broad Street pump revisited: response of volunteers to ingested cholera vibrios, *Bulletin of the New York Academy of Medicine*, **47**, 1181-1191.

Hoshino, K.*, et al.* (1998) Development and evaluation of a multiplex PCR assay for rapid detection of toxigenic *Vibrio cholerae* O1 and O139, *FEMS Immunology and Medical Microbiology*, **20**, 201-207.

Ichinose, Y.*, et al.* (1987) Enterotoxicity of El Tor-like hemolysin of non-O1 *Vibrio cholerae*, *Infection and Immunity*, **55**, 1090-1093.

Islam, M.S., Drasar, B.S. and Sack, R.B. (1994) Probable role of blue-green algae in maintaining endemicity and seasonality of cholera in Bangladesh: a hypothesis, *Journal of Diarrhoeal Diseases Research*, **12**, 245-256.

Jabeen, K., Zafar, A. and Hasan, R. (2008) Increased isolation of *Vibrio cholerae* O1 serotype Inaba over serotype Ogawa in Pakistan, *Eastern Mediterranean health journal = La revue de sante de la Mediterranee orientale = al-Majallah al-sihhiyah li-sharq al-mutawassit*, **14**, 564-570.

Jermyn, W.S. and Boyd, E.F. (2002) Characterization of a novel Vibrio pathogenicity island (VPI-2) encoding neuraminidase (nanH) among toxigenic *Vibrio cholerae* isolates, *Microbiology*, **148**, 3681-3693.

Kendall, E.A.*, et al.* (2010) Relatedness of *Vibrio cholerae* O1/O139 isolates from patients and their household contacts, determined by multilocus variable-number tandem-repeat analysis, *Journal of Bacteriology*, **192**, 4367-4376.

Kiiru, J.N.*, et al.* (2009) Molecular characterisation of *Vibrio cholerae* O1 strains carrying an SXT/R391-like element from cholera outbreaks in Kenya: 1994-2007, *BMC Microbiology*, **9**, 275.

Koch, R. (1884) An Address on Cholera and its Bacillus, *British medical journal*, **2**, 403-407.

Kondo, H.*, et al.* (2002) Post-flood--infectious diseases in Mozambique, *Prehospital and Disaster Medicine*, **17**, 126-133.

Koren, S.*, et al.* (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads, *Nature Biotechnology*, **30**, 693-700.

Korlach, J.*, et al.* (2010) Real-time DNA sequencing from single polymerase molecules, *Methods in Enzymology*, **472**, 431-455.

Kurtz, S.*, et al.* (2004) Versatile and open software for comparing large genomes, *Genome Biology*, **5**, R12.

Lam, C.*, et al.* (2010) Evolution of seventh cholera pandemic and origin of 1991 epidemic, Latin America, *Emerging Infectious Diseases*, **16**, 1130-1132.

Langmead, B.*, et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biology*, **10**, R25.

Langridge, G.C.*, et al.* (2009) Simultaneous assay of every Salmonella Typhi gene using one million transposon mutants, *Genome Research*, **19**, 2308-2316.

Larocque, R.C.*, et al.* (2005) Transcriptional profiling of *Vibrio cholerae* recovered directly from patient specimens during early and late stages of human infection, *Infection and Immunity*, **73**, 4488-4493.

Lee, S.H.*, et al.* (1999) Regulation and temporal expression patterns of *Vibrio cholerae* virulence genes during infection, *Cell*, **99**, 625-634.

Levine, M.M. (1997) Oral vaccines against cholera: lessons from Vietnam and elsewhere, *Lancet*, **349**, 220-221.

Levine, M.M.*, et al.* (1988) Volunteer studies of deletion mutants of *Vibrio cholerae* O1 prepared by recombinant techniques, *Infection and Immunity*, **56**, 161-167.

Li, H. and Durbin, R. (2010) Fast and accurate long-read alignment with Burrows-Wheeler transform, *Bioinformatics*, **26**, 589-595.

Li, H., Ruan, J. and Durbin, R. (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores, *Genome Research*, **18**, 1851-1858.

Li, R.*, et al.* (2010) De novo assembly of human genomes with massively parallel short read sequencing, *Genome Research*, **20**, 265-272.

Lindenbaum, J., Greenough, W.B. and Islam, M.R. (1967) Antibiotic therapy of cholera, *Bulletin of the World Health Organization*, **36**, 871-883.

Liu, L.*, et al.* (2012) Comparison of next-generation sequencing systems, *Journal of biomedicine & biotechnology*, **2012**, 251364.

Longini, I.M., Jr.*, et al.* (2002) Epidemic and endemic cholera trends over a 33-year period in Bangladesh, *Journal of Infectious Diseases*, **186**, 246-251.

Lopez, A.L.*, et al.* (2008) Cholera vaccines for the developing world, *Human Vaccines*, **4**, 165-169.

Mandal, S., Mandal, M.D. and Pal, N.K. (2011) Cholera: a great global concern, *Asian Pacific Journal of Tropical Medicine*, **4**, 573-580.

Manning, P.A. (1997) The tcp gene cluster of *Vibrio cholerae*, *Gene*, **192**, 63-70.

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

Merrell, D.S.*, et al.* (2002) Host-induced epidemic spread of the cholera bacterium, *Nature*, **417**, 642-645.

Miller, J.R., Koren, S. and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data, *Genomics*, **95**, 315-327.

Miotto, O.*, et al.* (2013) Multiple populations of artemisinin-resistant Plasmodium falciparum in Cambodia, *Nature Genetics*, **45**, 648-655.

Mohamed, A.A.*, et al.* (2012) Molecular epidemiology of geographically dispersed *Vibrio cholerae*, Kenya, January 2009-May 2010, *Emerging Infectious Diseases*, **18**, 925-931.

Mugoya, I.*, et al.* (2008) Rapid spread of *Vibrio cholerae* O1 throughout Kenya, 2005, *The American Journal of Tropical Medicine and Hygiene*, **78**, 527-533.

Mukhopadhyay, A.K.*, et al.* (1998) Emergence of fluoroquinolone resistance in strains of *Vibrio cholerae* isolated from hospitalized patients with acute diarrhea in Calcutta, India, *Antimicrobial Agents and Chemotherapy*, **42**, 206-207.

Mutreja, A.*, et al.* (2011) Evidence for several waves of global transmission in the seventh cholera pandemic, *Nature*, **477**, 462-465.

Mwansa, J.C.*, et al.* (2007) Multiply antibiotic-resistant *Vibrio cholerae* O1 biotype El Tor strains emerge during cholera outbreaks in Zambia, *Epidemiology and Infection*, **135**, 847-853.

Nair, G.B., Bhattacharya, S.K. and Deb, B.C. (1994) *Vibrio cholerae* O139 Bengal: the eighth pandemic strain of cholera, *Indian Journal of Public Health*, **38**, 33-36.

Nair, G.B.*, et al.* (2002) New variants of *Vibrio cholerae* O1 biotype El Tor with attributes of the classical biotype from hospitalized patients with acute diarrhea in Bangladesh, *Journal of Clinical Microbiology*, **40**, 3296-3299.

Nair, G.B.*, et al.* (2006) Cholera due to altered El Tor strains of *Vibrio cholerae* O1 in Bangladesh, *Journal of Clinical Microbiology*, **44**, 4211-4213.

Nelson, E.J.*, et al.* (2009) Cholera transmission: the host, pathogen and bacteriophage dynamic, *Nature Reviews Microbiology*, **7**, 693-702.

Nielsen, A.T.*, et al.* (2006) RpoS controls the *Vibrio cholerae* mucosal escape response, *PLoS Pathogens*, **2**, e109.

Ning, Z., Cox, A.J. and Mullikin, J.C. (2001) SSAHA: a fast search method for large DNA databases, *Genome Research*, **11**, 1725-1729.

Okoro, C.K., *et al.* (2012) Intracontinental spread of human invasive Salmonella Typhimurium pathovariants in sub-Saharan Africa, *Nature Genetics*, **44**, 1215-1221.

Pang, B., *et al.* (2007) Genetic diversity of toxigenic and nontoxigenic *Vibrio cholerae* serogroups O1 and O139 revealed by array-based comparative genomic hybridization, *Journal of Bacteriology*, **189**, 4837-4849.

Parsot, C. and Mekalanos, J.J. (1990) Expression of ToxR, the transcriptional activator of the virulence factors in *Vibrio cholerae*, is modulated by the heat shock response, *Proceedings of the National Academy of Sciences of the United States of America*, **87**, 9898-9902.

Pascual, M., *et al.* (2000) Cholera dynamics and El Nino-Southern Oscillation, *Science*, **289**, 1766-1769.

Peek, J.A. and Taylor, R.K. (1992) Characterization of a periplasmic thiol:disulfide interchange protein required for the functional maturation of secreted virulence factors of *Vibrio cholerae*, *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 6210-6214.

Perkins, T.T., *et al.* (2009) A strand-specific RNA-Seq analysis of the transcriptome of the typhoid bacillus Salmonella typhi, *PLoS Genetics*, **5**, e1000569.

Peterson, J.W., *et al.* (1979) Antigenic specificity of neutralizing antibody to cholera toxin, *Infection and Immunity*, **24**, 774-779.

Pierce, N.F. (1973) Differential inhibitory effects of cholera toxoids and ganglioside on the enterotoxins of *Vibrio cholerae* and Escherichia coli, *The Journal of Experimental Medicine*, **137**, 1009-1023.

Pop, M. and Salzberg, S.L. (2008) Bioinformatics challenges of new sequencing technology, *Trends in Genetics : TIG*, **24**, 142-149.

Quail, M.A., *et al.* (2008) A large genome center's improvements to the Illumina sequencing system, *Nature Methods*, **5**, 1005-1010.

Ramachandran, D., Bhanumathi, R. and Singh, D.V. (2007) Multiplex PCR for detection of antibiotic resistance genes and the SXT element: application in the characterization of *Vibrio cholerae*, *Journal of Medical Microbiology*, **56**, 346-351.

Ramamurthy, T., *et al.* (1993) Emergence of novel strain of *Vibrio cholerae* with epidemic potential in southern and eastern India, *Lancet*, **341**, 703-704.

Rashed, S.M., *et al.* (2012) Genetic characteristics of drug-resistant *Vibrio cholerae* O1 causing endemic cholera in Dhaka, 2006-2011, *Journal of Medical Microbiology*, **61**, 1736-1745.

Reimer, A.R., *et al.* (2011) Comparative genomics of *Vibrio cholerae* from Haiti, Asia, and Africa, *Emerging Infectious Diseases*, **17**, 2113-2121.

Roy, S.K., *et al.* (2008) Zinc supplementation in children with cholera in Bangladesh: randomised controlled trial, *BMJ*, **336**, 266-268.

Sack, D.A., *et al.* (1978) Single-dose doxycycline for cholera, *Antimicrobial Agents and Chemotherapy*, **14**, 462-464.

Sack, D.A., *et al.* (2004) Cholera, *Lancet*, **363**, 223-233.

Sack, D.A., *et al.* (1998) Validation of a volunteer model of cholera with frozen bacteria as the challenge, *Infection and Immunity*, **66**, 1968-1972.

Safa, A., Nair, G.B. and Kong, R.Y. (2010) Evolution of new variants of *Vibrio cholerae* O1, *Trends in Microbiology*, **18**, 46-54.

Sanger, F., *et al.* (1977) Nucleotide sequence of bacteriophage phi X174 DNA, *Nature*, **265**, 687-695.

Sanger, F. and Coulson, A.R. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase, *Journal of Molecular Biology*, **94**, 441-448.

Schild, S., *et al.* (2007) Genes induced late in infection increase fitness of *Vibrio cholerae* after release into the environment, *Cell Host & Microbe*, **2**, 264-277.

Scrascia, M., *et al.* (2006) Clonal relationship among *Vibrio cholerae* O1 El Tor strains causing the largest cholera epidemic in Kenya in the late 1990s, *Journal of Clinical Microbiology*, **44**, 3401-3404.

Scrascia, M., *et al.* (2009) Cholera in Ethiopia in the 1990 s: epidemiologic patterns, clonal analysis, and antimicrobial resistance, *International journal of Medical Microbiology : IJMM*, **299**, 367-372.

Scrascia, M., *et al.* (2009) Clonal relationship among *Vibrio cholerae* O1 El Tor strains isolated in Somalia, *International journal of medical microbiology : International Journal of Medical Microbiology*, **299**, 203-207.

Shapiro, R.L., *et al.* (1999) Transmission of epidemic *Vibrio cholerae* O1 in rural western Kenya associated with drinking water from Lake Victoria: an environmental reservoir for cholera?, *The American Journal of Tropical Medicine and hygiene*, **60**, 271-276.

Sharma, C.M., *et al.* (2010) The primary transcriptome of the major human pathogen Helicobacter pylori, *Nature*, **464**, 250-255.

Shendure, J. and Ji, H. (2008) Next-generation DNA sequencing, *Nature Biotechnology*, **26**, 1135-1145.

Shikanga, O.T., *et al.* (2009) High mortality in a cholera outbreak in western Kenya after post-election violence in 2008, *The American Journal of Tropical Medicine and Hygiene*, **81**, 1085-1090.

Silva, A.J., *et al.* (2006) Contribution of hemagglutinin/protease and motility to the pathogenesis of El Tor biotype cholera, *Infection and Immunity*, **74**, 2072-2079.

Sixma, T.K., *et al.* (1991) Crystal structure of a cholera toxin-related heat-labile enterotoxin from E. coli, *Nature*, **351**, 371-377.

Skorupski, K. and Taylor, R.K. (1997) Control of the ToxR virulence regulon in *Vibrio cholerae* by environmental stimuli, *Molecular Microbiology*, **25**, 1003-1009.

Stamatakis, A. (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models, *Bioinformatics*, **22**, 2688-2690.

Stroeher, U.H., *et al.* (1992) Serotype conversion in *Vibrio cholerae* O1, *Proceedings of the National Academy of Sciences of the United States of America*, **89**, 2566-2570.

Studholme, D.J., *et al.* (2009) A draft genome sequence and functional screen reveals the repertoire of type III secreted proteins of Pseudomonas syringae pathovar tabaci 11528, *BMC Genomics*, **10**, 395.

Sur, D., *et al.* (2009) Efficacy and safety of a modified killed-whole-cell oral cholera vaccine in India: an interim analysis of a cluster-randomised, double-blind, placebo-controlled trial, *Lancet*, **374**, 1694-1702.

Szabady, R.L., *et al.* (2011) TagA is a secreted protease of *Vibrio cholerae* that specifically cleaves mucin glycoproteins, *Microbiology*, **157**, 516-525.

Tacket, C.O., *et al.* (1998) Investigation of the roles of toxin-coregulated pili and mannose-sensitive hemagglutinin pili in the pathogenesis of *Vibrio cholerae* O139 infection, *Infection and Immunity*, **66**, 692-695.

Tanaka, S., *et al.* (2013) Transcriptome Analysis of Mouse Brain Infected with Toxoplasma gondii, *Infection and Immunity*.

Tauxe, R.V., Mintz, E.D. and Quick, R.E. (1995) Epidemic cholera in the new world: translating field epidemiology into new prevention strategies, *Emerging Infectious Diseases*, **1**, 141-146.

Taylor, R.K., *et al.* (1987) 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*, **84**, 2833-2837.

Trucksis, M.*, et al.* (2000) *Vibrio cholerae* ACE stimulates Ca(2+)-dependent Cl(-)/HCO(3)(-) secretion in T84 cells in vitro, *American Journal of Physiology. Cell Physiology*, **279**, C567-577.

van Opijnen, T. and Camilli, A. (2013) Transposon insertion sequencing: a new tool for systems-level analysis of microorganisms, *Nature Reviews Microbiology*, **11**, 435-442.

Waldor, M.K. and RayChaudhuri, D. (2000) Treasure trove for cholera research, *Nature*, **406**, 469-470.

Waldor, M.K., Tschape, H. and Mekalanos, J.J. (1996) A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* O139, *Journal of Bacteriology*, **178**, 4157-4165.

Watnick, P.I.*, et al.* (2001) The absence of a flagellum leads to altered colony morphology, biofilm development and virulence in *Vibrio cholerae* O139, *Mol Microbiology*, **39**, 223-235.

WHO (2011) Cholera, 2010, *Releve epidemiologique hebdomadaire / Section d'hygiene du Secretariat de la Societe des Nations = Weekly Epidemiological Record / Health Section of the Secretariat of the League of Nations*, **86**, 325-339.

Wozniak, R.A.*, et al.* (2009) Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs, *PLoS Genetics*, **5**, e1000786.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood, *Molecular Biology and Evolution*, **24**, 1586-1591.

Zerbino, D.R. and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs, *Genome Research*, **18**, 821-829.

Appendix

A.1: Table showing all the 1022 *V. cholerae* that were analysed in the expanded seventh pandemic phylogenetic study. 923 strains were sequenced at the WTSI and the rest were published or publicly available genomes for which the reference has been provided. The accession numbers and metadata for each isolate is also provided.

Isolate	Source	Time	ENA Accession
4605	India	2007	ERS013257
4656	India	2006	ERS013258
4675	Bangladesh	2001	ERS013259
4679	Bangladesh	1999	ERS013260
4663	Bangladesh	2001	ERS013261
4661	Bangladesh	2001	ERS013263
4660	Bangladesh	1994	ERS013262
4110	Vietnam	1995	ERS013252
4111	Vietnam	2002	ERS013253
4322	India	2004	ERS013254
4642	India	2006	ERS013255
4670	Bangladesh	1991	ERS013256
4672	Bangladesh	2000	ERS016137
4122	Vietnam	2007	ERS013264
BangladeshA330	India	1993	ERS013124
IndiaMBN17	India	2004	ERS013134
IndiaVC51	India	1992	ERS013133
IndiaV212_1	India	1991	ERS013132
IndiaV109	India	1990	ERS013131
IndiaV5	India	1989	ERS013130
BangladeshA488	Bangladesh	2006	ERS013129
IndiaMBRN14	India	2004	ERS013127
IndiaMJ1485	Bangladesh	1994	ERS013126
BangladeshA383	Bangladesh	2002	ERS013125
IndiaMG160	Bangladesh	1991	ERS013135
KoreaA177	Colombia	1992	ERS013182
KoreaA155	Mozambique	1991	ERS013181
KoreaA154	Mozambique	1991	ERS013180
KoreaA152	Mozambique	1991	ERS013179
KoreaA131	India	1989	ERS013178
KoreaA130	India	1989	ERS013177
KoreaA185	Colombia	1992	ERS013175
KoreaA184	Colombia	1992	ERS013174

KoreaA109	N.I	1990	ERS013173
KoreaA180	Colombia	1992	ERS013183
KoreaA186	Argentina	1992	ERS013184
KoreaA201	Argentina	1992	ERS013189
KoreaA200	Argentina	1992	ERS013188
KoreaA241	Vietnam	1989	ERS013187
KoreaA232	Mexico	1991	ERS013186
KoreaA193	Bolivia	1992	ERS013185
KoreaA231	Mexico	1991	ERS013195
KoreaA245	Vietnam	1989	ERS013196
KoreaA481	Djibouti	2007	ERS013206
KoreaA397	Bangladesh	1987	ERS013205
KoreaA390	Argentina	1993	ERS013204
KoreaA346	Bangladesh	1994	ERS013202
KoreaA316	Bangladesh	1987	ERS013200
KoreaA487	Bangladesh	2007	ERS013199
KoreaA483	Djibouti	2007	ERS013198
KoreaA482	Djibouti	2007	ERS013207
IndiaGP152	India	1979	ERS013146
IndiaPRL5	India	1979	ERS013145
IndiaGP145	India	1979	ERS013143
IndiaGP143	Bahrain	1978	ERS013142
IndiaGP140	Malaysia	1978	ERS013141
IndiaGP106	W.Germany	1975	ERS013140
IndiaPRL64	India	1992	ERS013139
IndiaPRL18	India	1984	ERS013138
IndiaGP60	India	1973	ERS013137
IndiaIDHO1_726	India	2009	ERS013147
Kenya6180	Kenya	2007	ERS013208
Kenya6210	Kenya	2007	ERS013218
Kenya6201	Kenya	2007	ERS013217
Kenya6197	Kenya	2007	ERS013216
Kenya6196	Kenya	2005	ERS013215
Kenya6195	Kenya	2005	ERS013214
Kenya6194	Kenya	2007	ERS013213
Kenya6193	Kenya	2005	ERS013212
Kenya6215	Kenya	2005	ERS013211
Kenya6214	Kenya	2007	ERS013210
Kenya6191	Kenya	2005	ERS013209
Kenya6212	Kenya	2007	ERS013219
Kenya7682	Kenya	2009	ERS013220
Kenya7687	Kenya	2009	ERS013226
Kenya7686	Kenya	2009	ERS013225
Kenya7685	Kenya	2009	ERS013224

Kenya7684	Kenya	2009	ERS013221
1346	Mozambique	2005	ERS013265
4551	India	2007	ERS013266
4623	India	2007	ERS013267
4593	India	2007	ERS013268
4538	India	2007	ERS013269
4339	India	2004	ERS013270
4121	Vietnam	2004	ERS013271
4113	Vietnam	2003	ERS013273
4585	India	2007	ERS013232
4552	India	2007	ERS013233
4488	India	2006	ERS013234
4784	Tanzania	2009	ERS013235
4600	India	2007	ERS013236
4646	India	2007	ERS013237
4662	Bangladesh	2001	ERS013238
4519	India	2005	ERS013239
4536	India	2007	ERS013240
1326	Mozambique	2005	ERS013241
1672	Mozambique	2005	ERS013242
4698	India	1980	ERS013243
4713	NI	1973	ERS013244
4714	Angola	1989	ERS013245
4715	Indonesia	1957	ERS013246
4716	Bangladesh	1979	ERS013247
4717	India	1977	ERS013248
4718	Bangladesh	1975	ERS013249
4719	Bangladesh	1979	ERS013250
4720	Peru	1991	ERS013251
4721	Peru	1991	ERS013274
4722	Peru	1991	ERS013275
4723	Peru	1991	ERS013276
4764	Bangladesh	1994	ERS013278
4772	Bangladesh	N.I	ERS013279
4771	Bangladesh	2007	ERS013281
MG16226	Bangladesh	1991	ERS013282
V6	Kolkata	2006	ERS025559
V7	Kolkata	2006	ERS025560
V8	Kolkata	2006	ERS025561
V11	Kolkata	2006	ERS025562
V12	Kolkata	2006	ERS025563
V13	Kolkata	2006	ERS025564
V14	Kolkata	2006	ERS025565
V15	Kolkata	2006	ERS025566

V16	Kolkata	2006	ERS025567
V17	Kolkata	2006	ERS025568
V29	Kolkata	2007	ERS025569
V30	Kolkata	2007	ERS025570
V31	Kolkata	2007	ERS025571
V32	Kolkata	2007	ERS025572
V33	Kolkata	2007	ERS025573
V35	Kolkata	2007	ERS025574
V36	Kolkata	2007	ERS025575
V46	Kolkata	2007	ERS025576
V47	Kolkata	2007	ERS025577
V49	Kolkata	2007	ERS025578
V92	Kolkata	2007	ERS025579
V93	Kolkata	2007	ERS025580
V94	Kolkata	2007	ERS025581
V95	Kolkata	2007	ERS025582
V96	Kolkata	2007	ERS025583
V97	Kolkata	2007	ERS025584
V98	Kolkata	2007	ERS025585
V99	Kolkata	2007	ERS025586
V100	Kolkata	2007	ERS025587
V101	Kolkata	2007	ERS025588
V102	Kolkata	2007	ERS025589
V103	Kolkata	2007	ERS025590
V104	Kolkata	2007	ERS025591
V105	Kolkata	2007	ERS025592
V107	Kolkata	2007	ERS025593
V108	Kolkata	2007	ERS025594
V109	Kolkata	2007	ERS025595
V110	Kolkata	2007	ERS025596
V111	Kolkata	2007	ERS025597
V113	Kolkata	2007	ERS025598
V114	Kolkata	2007	ERS025599
IB5234	Haiti	2010	ERS031632
IB5236	Haiti	2010	ERS031633
IB5240	Haiti	2010	ERS031634
IB5241	Haiti	2010	ERS031635
IB5261	Haiti	2010	ERS031636
IB5275	Nepal	2007	ERS031637
IB5276	Nepal	2007	ERS031638
IB5277	Nepal	2007	ERS031639
IB5278	Nepal	2007	ERS031640
IB5279	Nepal	2007	ERS031641
IB5280	Nepal	2007	ERS031642

IB5281	Nepal	2007	ERS031643
IB5282	Nepal	2007	ERS031644
IB5283	Nepal	2007	ERS031645
IB5284	Nepal	2007	ERS031646
IB5285	Nepal	2007	ERS031647
IB5286	Nepal	2007	ERS031648
IB5287	Nepal	2007	ERS031649
IB5288	Nepal	2007	ERS031650
IB5289	Nepal	2007	ERS031651
IB5290	Nepal	2007	ERS031652
IB5291	Nepal	2007	ERS031653
IB5292	Nepal	2007	ERS031654
IB5293	Nepal	2007	ERS031655
IB5294	Nepal	2007	ERS031656
IB5295	Nepal	2008	ERS031657
IB5296	Nepal	2008	ERS031658
IB5297	Nepal	2008	ERS031659
IB5298	Nepal	2008	ERS031660
IB5299	Nepal	2008	ERS031661
IB5300	Nepal	2008	ERS031662
IB5301	Nepal	2008	ERS031663
IB5302	Nepal	2008	ERS031664
IB5303	Nepal	2008	ERS031665
IB5304	Nepal	2008	ERS031666
IB5305	Nepal	2008	ERS031667
IB5306	Nepal	2008	ERS031668
IB5307	Nepal	2008	ERS031669
IB5308	Nepal	2008	ERS031670
IB5309	Nepal	2008	ERS031671
IB5310	Nepal	2008	ERS031672
IB5311	Nepal	2008	ERS031673
IB5312	Nepal	2008	ERS031674
IB5313	Nepal	2008	ERS031675
IB5314	Nepal	2008	ERS031676
IB5315	Nepal	2009	ERS031677
IB5316	Nepal	2009	ERS031678
IB5317	Nepal	2009	ERS031679
IB5318	Nepal	2009	ERS031680
IB5319	Nepal	2009	ERS031681
IB5320	Nepal	2009	ERS031682
IB5321	Nepal	2009	ERS031683
IB5322	Nepal	2009	ERS031684
IB5323	Nepal	2009	ERS031685
IB5324	Nepal	2009	ERS031686

IB5325	Nepal	2010	ERS031687
IB5326	Nepal	2010	ERS031688
IB5327	Nepal	2010	ERS031689
IB5328	Nepal	2010	ERS031690
IB5329	Nepal	2010	ERS031691
IB5330	Nepal	2010	ERS031692
IB5331	Nepal	2010	ERS031693
IB5332	Nepal	2010	ERS031694
IB5333	Nepal	2010	ERS031695
IB5334	Nepal	2010	ERS031696
Gaz2	Gaza	1994	ERS032744
Gaz4	Gaza	1994	ERS032745
Gaz6	Gaza	1994	ERS032746
Gaz8	Gaza	1994	ERS032747
Gaz11	Gaza	1994	ERS032748
Gaz13	Gaza	1994	ERS032749
Gaz14	Gaza	1994	ERS032750
Gaz17	Gaza	1994	ERS032751
Gaz18	Gaza	1994	ERS032752
Gaz21	Gaza	1994	ERS032753
Gaz23	Gaza	1994	ERS032754
Gaz26	Gaza	1994	ERS032755
Gaz27	Gaza	1994	ERS032756
Gaz28	Gaza	1994	ERS032757
Gaz33	Gaza	1994	ERS032758
Gaz34	Gaza	1994	ERS032759
1_1970	Jerusalem	1970	ERS032760
2_1970	Jerusalem	1970	ERS032761
3_1970	Jerusalem	1970	ERS032762
11_1970	Jerusalem	1970	ERS032763
12_1970	Jerusalem	1970	ERS032764
1100_1970	Jerusalem	1970	ERS032766
101_72	Jerusalem	1970	ERS032767
Mex6	Mexico	1992	ERS032769
IB4513	Kolkata	2006	ERS032772
IB4515	Kolkata	2006	ERS032773
IB4517	Kolkata	2006	ERS032774
IB4518	Kolkata	2006	ERS032775
IB4520	Kolkata	2006	ERS032776
IB4526	Kolkata	2006	ERS032777
IB4527	Kolkata	2006	ERS032778
IB4535	Kolkata	2006	ERS032779
IB4536	Kolkata	2007	ERS032780
IB4537	Kolkata	2007	ERS032781

IB4538	Kolkata	2007	ERS032782
IB4539	Kolkata	2007	ERS032783
IB4540	Kolkata	2007	ERS032784
IB4541	Kolkata	2007	ERS032785
IB4542	Kolkata	2007	ERS032786
IB4543	Kolkata	2007	ERS032787
IB4545	Kolkata	2007	ERS032788
IB4546	Kolkata	2007	ERS032789
IB4580	Kolkata	2007	ERS032812
IB4581	Kolkata	2007	ERS032813
IB4582	Kolkata	2007	ERS032814
IB4583	Kolkata	2007	ERS032815
IB4584	Kolkata	2007	ERS032816
IB4585	Kolkata	2007	ERS032817
IB4586	Kolkata	2007	ERS032818
IB4587	Kolkata	2007	ERS032819
IB4588	Kolkata	2007	ERS032820
IB4589	Kolkata	2007	ERS032821
IB4590	Kolkata	2007	ERS032822
IB4591	Kolkata	2007	ERS032823
IB4592	Kolkata	2007	ERS032824
IB4593	Kolkata	2007	ERS032825
IB4594	Kolkata	2007	ERS032826
IB4595	Kolkata	2007	ERS032827
IB4596	Kolkata	2007	ERS032828
IB4597	Kolkata	2007	ERS032829
IB4598	Kolkata	2007	ERS032830
IB4599	Kolkata	2007	ERS032831
IB4600	Kolkata	2007	ERS032832
IB4601	Kolkata	2007	ERS032833
IB4602	Kolkata	2007	ERS032834
IB4603	Kolkata	2007	ERS032835
IB4604	Kolkata	2007	ERS032836
IB4605	Kolkata	2007	ERS032837
IB4606	Kolkata	2007	ERS032838
IB4607	Kolkata	2007	ERS032839
IB4608	Kolkata	2007	ERS032840
IB4609	Kolkata	2007	ERS032841
IB4610	Kolkata	2007	ERS032842
IB4626	Kolkata	2007	ERS032843
IB4647	Kolkata	2007	ERS032844
116072	Mexico	1991	ERS027692
87258	Mexico	1991	ERS027693
116073	Mexico	1991	ERS027694

87151	Mexico	1992	ERS027695
116075	Mexico	1992	ERS027696
87397	Mexico	1993	ERS027697
87667	Mexico	1993	ERS027698
87662	Mexico	1993	ERS027699
87406	Mexico	1994	ERS027700
87409	Mexico	1995	ERS027701
95430	Mexico	1997	ERS027704
354_02	Zambia	1996	ERS027707
204_12	Zambia	2003	ERS027708
B-33	Mozambique	2004	ERS027709
B-64	Mozambique	2004	ERS027710
Zim-12	Zimbabwe	2009	ERS027711
Zim-25	Zimbabwe	2009	ERS027712
Zim-27	Zimbabwe	2009	ERS027713
IB4553	Kolkata	2007	ERS032790
IB4556	Kolkata	2007	ERS032791
IB4557	Kolkata	2007	ERS032792
IB4558	Kolkata	2007	ERS032793
IB4559	Kolkata	2007	ERS032794
IB4560	Kolkata	2007	ERS032795
IB4561	Kolkata	2007	ERS032796
IB4562	Kolkata	2007	ERS032797
IB4563	Kolkata	2007	ERS032798
IB4564	Kolkata	2007	ERS032799
IB4567	Kolkata	2007	ERS032800
IB4569	Kolkata	2007	ERS032801
IB4570	Kolkata	2007	ERS032802
IB4571	Kolkata	2007	ERS032803
IB4572	Kolkata	2007	ERS032804
IB4573	Kolkata	2007	ERS032805
IB4574	Kolkata	2007	ERS032806
IB4575	Kolkata	2007	ERS032807
IB4576	Kolkata	2007	ERS032808
IB4577	Kolkata	2007	ERS032809
IB4578	Kolkata	2007	ERS032810
IB4579	Kolkata	2007	ERS032811
F1DN4	Pakistan	2010	ERS032845
F2D59	Pakistan	2010	ERS032846
F4D48	Pakistan	2010	ERS032848
F5D38	Pakistan	2010	ERS032849
F7D30	Pakistan	2010	ERS032851
F8D25	Pakistan	2010	ERS032852
F11D4	Pakistan	2010	ERS032855

F12D1	Pakistan	2010	ERS032856
F14KPD3	Pakistan	2010	ERS032858
F15KTH7	Pakistan	2010	ERS032859
F16KTH6	Pakistan	2010	ERS032860
F17KTH4	Pakistan	2010	ERS032861
F18KTH3	Pakistan	2010	ERS032862
F19KTH2	Pakistan	2010	ERS032863
S1KCH15	Pakistan	2010	ERS032864
S2KCH17	Pakistan	2010	ERS032865
S4KCH16	Pakistan	2010	ERS032867
S5KCH10	Pakistan	2010	ERS032868
S6KCH7	Pakistan	2010	ERS032869
S7KCH20	Pakistan	2010	ERS032870
S8KCH18	Pakistan	2010	ERS032871
S9KCH9	Pakistan	2010	ERS032872
S10P57	Pakistan	2010	ERS032873
S12P76	Pakistan	2010	ERS032875
S13P83	Pakistan	2010	ERS032876
S14P9	Pakistan	2010	ERS032877
S16HH1	Pakistan	2010	ERS032879
S17HH3	Pakistan	2010	ERS032880
S18HH4	Pakistan	2010	ERS032881
S19HH5	Pakistan	2010	ERS032882
S20HH14	Pakistan	2010	ERS032883
S21HH15	Pakistan	2010	ERS032884
S22HH17	Pakistan	2010	ERS032885
S23HH18	Pakistan	2010	ERS032886
S24RG6	Pakistan	2010	ERS032887
S25R22	Pakistan	2010	ERS032888
S26R24	Pakistan	2010	ERS032889
S27RG11	Pakistan	2010	ERS032890
B33	Mozambique	2004	Chun et al 2009, PNAS, 106(36): 15442-7
CIRS101	Bangladesh	2002	Chun et al 2009, PNAS, 106(36): 15442-7
MJ1236	Bangladesh	1994	Chun et al 2009, PNAS, 106(36): 15442-7
MO10	India	1992	Chun et al 2009, PNAS, 106(36): 15442-7
RC9	Kenya	1985	Chun et al 2009, PNAS, 106(36): 15442-7
2010EL_1786	Haiti	2010	Chin et al 2011, NEJM, 364(1): 33-42
2010EL_1792	Haiti	2010	Chin et al 2011, NEJM, 364(1): 33-42
2010EL_1798	Haiti	2010	Chin et al 2011, NEJM, 364(1): 33-42
6734_09	India	2009	NA
6801_09	India	2009	NA
7683_09	India	2009	NA
7772_09	India	2009	NA
7934_09	India	2009	NA

7994_09	India	2009	NA
9088_09	India	2009	NA
6554_09	India	2009	NA
6557_09	India	2009	NA
6702_09	India	2009	NA
6361_09	India	2009	NA
6394_09	India	2009	NA
6099_09	India	2009	NA
6111_09	India	2009	NA
6235_09	India	2009	NA
6236_09	India	2009	NA
6259_09	India	2009	NA
5185_09	India	2009	NA
5186_09	India	2009	NA
5188_09	India	2009	NA
5189_09	India	2009	NA
5197_09	India	2009	NA
5198_09	India	2009	NA
5202_09	India	2009	NA
5235_09	India	2009	NA
5238_09	India	2009	NA
5286_09	India	2009	NA
5365_09	India	2009	NA
5366_09	India	2009	NA
5417_09	India	2009	NA
5663_09	India	2009	NA
6016_09	India	2009	NA
6019_09	India	2009	NA
4966_09	India	2009	NA
5046_09	India	2009	NA
5064_09	India	2009	NA
5157_09	India	2009	NA
5159_09	India	2009	NA
7929_1991	Mexico	1991	ERS135702
8012_1991	Mexico	1991	ERS135834
8022_1991	Mexico	1991	ERS135704
8204_1991	Mexico	1991	ERS135705
8338_1991	Mexico	1991	ERS136087
16974_1992	Mexico	1992	ERS135707
19294_1992	Mexico	1992	ERS135708
33297_1993	Mexico	1993	ERS135836
54267_1994	Mexico	1994	ERS135710
54328_1994	Mexico	1994	ERS135711
60452_1995	Mexico	1995	ERS135837

60483_1995	Mexico	1995	ERS135713
1401_2004	Mexico	2004	ERS135714
1992_2004	Mexico	2004	ERS136025
2006_2004	Mexico	2004	ERS135716
2007_2004	Mexico	2004	ERS135717
985_2007	Mexico	2007	ERS135839
2533_2007	Mexico	2007	ERS135719
353_2008	Mexico	2008	ERS135840
354_2008	Mexico	2008	ERS135722
372_2008	Mexico	2008	ERS135723
504_2008	Mexico	2008	ERS136026
971_2008	Mexico	2008	ERS135725
210_2010	Mexico	2010	ERS135728
211_2010	Mexico	2010	ERS135729
601_2010	Mexico	2010	ERS135732
2496_2010	Mexico	2010	ERS135737
204_2010	Mexico	2010	ERS135740
S1PS7	Pakistan	2011	ERS135741
S2PS18	Pakistan	2011	ERS136028
S3PS25	Pakistan	2011	ERS135743
S4769	Pakistan	2011	ERS135744
S5N5	Pakistan	2011	ERS135848
S6N7	Pakistan	2011	ERS135746
S7N10	Pakistan	2011	ERS135747
S8NP3	Pakistan	2011	ERS135849
S9NP5	Pakistan	2011	ERS135749
S10NP6	Pakistan	2011	ERS135750
S11NP7	Pakistan	2011	ERS136029
S12NP14	Pakistan	2011	ERS135752
S13CS1	Pakistan	2011	ERS135753
S14CS12	Pakistan	2011	ERS135851
S15CS15	Pakistan	2011	ERS135755
S16CS16	Pakistan	2011	ERS135756
S17CS18	Pakistan	2011	ERS135852
S18770	Pakistan	2011	ERS135758
S19751	Pakistan	2011	ERS135759
S20759	Pakistan	2011	ERS136108
S21760	Pakistan	2011	ERS135761
S22754	Pakistan	2011	ERS135762
S23756	Pakistan	2011	ERS135854
S24758	Pakistan	2011	ERS135764
S25763	Pakistan	2011	ERS135765
S26753	Pakistan	2011	ERS135855
S27750	Pakistan	2011	ERS135767

S28703	Pakistan	2011	ERS135768
S29709	Pakistan	2011	ERS136031
S30719	Pakistan	2011	ERS135770
S31722	Pakistan	2011	ERS135771
S32729	Pakistan	2011	ERS135857
S33732	Pakistan	2011	ERS135773
S34736	Pakistan	2011	ERS135774
S35739	Pakistan	2011	ERS135858
S36742	Pakistan	2011	ERS135776
S37F5	Pakistan	2011	ERS135777
S38F6	Pakistan	2011	ERS136032
S39764	Pakistan	2011	ERS135779
S40767	Pakistan	2011	ERS135780
S41768	Pakistan	2011	ERS135860
S42774	Pakistan	2011	ERS135782
S43A4	Pakistan	2011	ERS135783
S44755	Pakistan	2011	ERS135861
S45757	Pakistan	2011	ERS135785
S46765	Pakistan	2011	ERS135786
S47761	Pakistan	2011	ERS136090
S48BW5	Pakistan	2011	ERS135788
S49773	Pakistan	2011	ERS135789
S50771	Pakistan	2011	ERS135863
S51772	Pakistan	2011	ERS135791
S52776	Pakistan	2011	ERS135792
S53775	Pakistan	2011	ERS135864
S54762	Pakistan	2011	ERS135794
S55GB39	Pakistan	2011	ERS135795
S56BH11	Pakistan	2011	ERS136034
S57BH20	Pakistan	2011	ERS135797
S58BHJ	Pakistan	2011	ERS135798
S59BHA	Pakistan	2011	ERS135866
S60752	Pakistan	2011	ERS135800
A034_Vc	NA	1970	ERS135867
A328_Vc	NA	1993	ERS135807
A351_Vc	NA	1994	ERS135869
A352_Vc	NA	1995	ERS135809
A356_Vc	NA	1996	ERS135810
A359_Vc	NA	2002	ERS135870
A363_Vc	NA	2002	ERS135812
280_02	Zambia	1996	ERS136091
329_02	Zambia	1996	ERS135815
177_03	Zambia	1997	ERS135816
330_02	Zambia	1997	ERS135872

236_02	Zambia	1997	ERS135818
202_02	Zambia	1997	ERS135819
20_03	Zambia	1997	ERS135873
143_12	Zambia	2003	ERS135821
169_12	Zambia	2003	ERS135822
218_02	Zambia	2003	ERS136037
224_12	Zambia	2003	ERS135824
329_01	Zambia	2004	ERS135825
276_01	Zambia	2004	ERS135875
260_01	Zambia	2004	ERS135827
178_02	Zambia	2004	ERS135828
187_03	Zambia	2004	ERS135876
259_01	Zambia	2004	ERS135830
179_02	Zambia	2004	ERS135831
192_02	Zambia	2004	ERS136038
336_01	Zambia	2004	ERS135833
IB5398	Nepal	2010	ERS044944
IB5399	Nepal	2010	ERS044945
IB5400	Nepal	2010	ERS044946
IB5401	Nepal	2010	ERS044947
IB5402	Nepal	2010	ERS044948
IB5403	Nepal	2010	ERS044949
IB5404	Nepal	2010	ERS044950
IB5405	Nepal	2010	ERS044951
IB5406	Nepal	2010	ERS044952
IB5407	Nepal	2010	ERS044953
IB5408	Nepal	2010	ERS044954
IB5409	Nepal	2010	ERS044955
IB5410	Nepal	2010	ERS044956
IB5411	Nepal	2010	ERS044957
IB5412	Nepal	2010	ERS044958
IB5413	Nepal	2010	ERS044959
IB5414	Nepal	2010	ERS044960
IB5415	Nepal	2010	ERS044961
IB5416	Nepal	2010	ERS044962
IB5417	Nepal	2010	ERS044963
IB5418	Nepal	2010	ERS044964
IB5419	Nepal	2010	ERS044965
IB5420	Nepal	2010	ERS044966
IB5421	Nepal	2010	ERS044967
IB5422	Nepal	2010	ERS044968
IB5423	Nepal	2010	ERS044969
IB5424	Nepal	2010	ERS044970
N16961M1	Bangladesh	2011	ERS051517

N16961M2	Bangladesh	2011	ERS051518
N16961M3	Bangladesh	2011	ERS051519
N16961M4	Bangladesh	2011	ERS051520
PSVC1	Bangladesh	2009	ERS051521
PRVC2	Bangladesh	2009	ERS051522
PSVC1M1	Bangladesh	2011	ERS051523
PSVC1M2	Bangladesh	2011	ERS051524
PSVC1M3	Bangladesh	2011	ERS051525
PSVC1M4	Bangladesh	2011	ERS051526
Prevac4338	Kolkata	2004	ERS070649
Prevac4339	Kolkata	2004	ERS070650
Prevac4341	Kolkata	2004	ERS070651
Prevac4342	Kolkata	2004	ERS070652
Prevac4344	Kolkata	2004	ERS070653
Prevac4345	Kolkata	2004	ERS070654
Prevac4346	Kolkata	2004	ERS070655
Prevac4347	Kolkata	2004	ERS070656
Prevac4348	Kolkata	2004	ERS070657
Prevac4351	Kolkata	2004	ERS070658
Prevac4352	Kolkata	2004	ERS070659
Prevac4353	Kolkata	2004	ERS070660
Prevac4354	Kolkata	2004	ERS070661
Prevac4355	Kolkata	2004	ERS070662
Prevac4356	Kolkata	2004	ERS070663
Prevac4357	Kolkata	2004	ERS070664
Prevac4358	Kolkata	2004	ERS070665
Prevac4359	Kolkata	2004	ERS070666
Prevac4361	Kolkata	2004	ERS070667
Prevac4362	Kolkata	2004	ERS070668
Prevac4363	Kolkata	2004	ERS070669
Prevac4364	Kolkata	2004	ERS070670
Prevac4365	Kolkata	2004	ERS070671
Prevac4368	Kolkata	2004	ERS070673
Prevac4369	Kolkata	2004	ERS070674
Prevac4370	Kolkata	2004	ERS070675
Prevac4371	Kolkata	2004	ERS070676
Prevac4372	Kolkata	2004	ERS070677
Prevac4373	Kolkata	2004	ERS070678
Prevac4374	Kolkata	2004	ERS070679
Prevac4375	Kolkata	2004	ERS070680
Prevac4376	Kolkata	2004	ERS070681
Prevac4377	Kolkata	2004	ERS070682
Prevac4378	Kolkata	2004	ERS070683
Prevac4379	Kolkata	2004	ERS070684

Prevac4380	Kolkata	2004	ERS070685
Prevac4381	Kolkata	2004	ERS070686
Prevac4382	Kolkata	2004	ERS070687
Prevac4383	Kolkata	2004	ERS070688
Prevac4384	Kolkata	2004	ERS070689
Prevac4343	Kolkata	2004	ERS070690
Prevac4386	Kolkata	2004	ERS070691
Prevac4387	Kolkata	2004	ERS070692
Prevac4388	Kolkata	2004	ERS070693
Prevac4389	Kolkata	2004	ERS070694
Prevac4390	Kolkata	2004	ERS070695
Prevac4391	Kolkata	2004	ERS070696
Prevac4392	Kolkata	2004	ERS070697
Prevac4393	Kolkata	2004	ERS070698
Prevac4394	Kolkata	2004	ERS070699
Prevac4395	Kolkata	2004	ERS070700
Prevac4396	Kolkata	2004	ERS070701
Prevac4397	Kolkata	2004	ERS070702
Prevac4398	Kolkata	2004	ERS070703
Prevac4399	Kolkata	2004	ERS070704
Prevac4400	Kolkata	2004	ERS070705
Prevac4401	Kolkata	2004	ERS070706
Prevac4403	Kolkata	2004	ERS070707
Prevac4404	Kolkata	2005	ERS070708
Prevac4405	Kolkata	2005	ERS070709
Prevac4406	Kolkata	2005	ERS070710
Prevac4407	Kolkata	2005	ERS070711
Prevac4408	Kolkata	2005	ERS070712
Prevac4409	Kolkata	2005	ERS070713
Prevac4410	Kolkata	2005	ERS070714
Prevac4411	Kolkata	2005	ERS070715
Prevac4412	Kolkata	2005	ERS070716
Prevac4413	Kolkata	2005	ERS070717
Prevac4414	Kolkata	2005	ERS070718
Prevac4415	Kolkata	2005	ERS070719
Prevac4416	Kolkata	2005	ERS070720
Prevac4417	Kolkata	2005	ERS070721
Prevac4418	Kolkata	2005	ERS070722
Prevac4419	Kolkata	2005	ERS070723
Prevac4420	Kolkata	2005	ERS070724
Prevac4421	Kolkata	2005	ERS070725
Prevac4422	Kolkata	2005	ERS070726
Prevac4423	Kolkata	2005	ERS070727
Prevac4424	Kolkata	2005	ERS070728

Prevac4425	Kolkata	2005	ERS070729
Prevac4426	Kolkata	2005	ERS070730
Prevac4427	Kolkata	2005	ERS070731
Prevac4428	Kolkata	2005	ERS070732
Prevac4429	Kolkata	2006	ERS070733
Prevac4430	Kolkata	2006	ERS070734
Prevac4431	Kolkata	2006	ERS070735
Prevac4433	Kolkata	2006	ERS070737
Prevac4434	Kolkata	2006	ERS070738
Prevac4435	Kolkata	2006	ERS070739
Prevac4436	Kolkata	2006	ERS070740
Prevac4437	Kolkata	2006	ERS070741
Prevac4438	Kolkata	2006	ERS070742
Prevac4439	Kolkata	2006	ERS070743
Prevac4440	Kolkata	2006	ERS070744
Prevac4441	Kolkata	2006	ERS070745
Prevac4442	Kolkata	2006	ERS070746
Prevac4443	Kolkata	2006	ERS070747
Prevac4444	Kolkata	2006	ERS070748
Prevac4445	Kolkata	2006	ERS070749
Prevac4446	Kolkata	2006	ERS070750
Prevac4447	Kolkata	2006	ERS070751
Prevac4448	Kolkata	2006	ERS070752
Prevac4449	Kolkata	2006	ERS070753
Prevac4450	Kolkata	2006	ERS070754
Prevac4451	Kolkata	2006	ERS070755
Prevac4452	Kolkata	2006	ERS070756
Prevac4453	Kolkata	2006	ERS070757
Prevac4454	Kolkata	2006	ERS070758
Prevac4455	Kolkata	2006	ERS070759
Prevac4456	Kolkata	2006	ERS070760
Prevac4457	Kolkata	2006	ERS070761
Prevac4458	Kolkata	2006	ERS070762
Prevac4459	Kolkata	2006	ERS070763
Prevac4460	Kolkata	2006	ERS070764
Prevac4461	Kolkata	2006	ERS070765
Prevac4462	Kolkata	2006	ERS070766
Prevac4463	Kolkata	2006	ERS070767
Prevac4464	Kolkata	2006	ERS070768
Prevac4465	Kolkata	2006	ERS070769
Prevac4466	Kolkata	2006	ERS070770
Prevac4467	Kolkata	2006	ERS070771
Prevac4468	Kolkata	2006	ERS070772
Prevac4469	Kolkata	2006	ERS070773

Prevac4470	Kolkata	2006	ERS070774
Prevac4471	Kolkata	2006	ERS070775
Prevac4472	Kolkata	2006	ERS070776
Prevac4473	Kolkata	2006	ERS070777
Prevac4474	Kolkata	2006	ERS070778
Prevac4475	Kolkata	2006	ERS070779
Prevac4476	Kolkata	2006	ERS070780
Prevac4477	Kolkata	2006	ERS070781
Prevac4478	Kolkata	2006	ERS070782
Prevac4479	Kolkata	2006	ERS070783
Prevac4480	Kolkata	2006	ERS070784
Prevac4481	Kolkata	2006	ERS070785
Prevac4482	Kolkata	2006	ERS070786
Prevac4483	Kolkata	2006	ERS070787
Prevac4484	Kolkata	2006	ERS070788
Prevac4485	Kolkata	2006	ERS070789
Prevac4486	Kolkata	2006	ERS070790
Prevac4487	Kolkata	2006	ERS070791
Prevac4488	Kolkata	2006	ERS070792
Prevac4489	Kolkata	2006	ERS070793
Prevac4490	Kolkata	2006	ERS070794
Prevac4491	Kolkata	2006	ERS070795
Prevac4492	Kolkata	2006	ERS070796
Prevac4493	Kolkata	2006	ERS070797
Prevac4495	Kolkata	2006	ERS070799
Prevac4496	Kolkata	2006	ERS070800
Prevac4497	Kolkata	2006	ERS070801
Prevac4498	Kolkata	2006	ERS070802
Prevac4500	Kolkata	2006	ERS070803
Prevac4501	Kolkata	2006	ERS070804
Prevac4502	Kolkata	2006	ERS070805
Prevac4503	Kolkata	2006	ERS070806
Prevac4504	Kolkata	2006	ERS070807
Prevac4505	Kolkata	2006	ERS070808
Prevac4506	Kolkata	2006	ERS070809
Prevac4507	Kolkata	2006	ERS070810
Prevac4508	Kolkata	2006	ERS070811
Prevac4509	Kolkata	2006	ERS070812
Prevac4510	Kolkata	2006	ERS070813
Prevac4511	Kolkata	2006	ERS070814
Prevac4512	Kolkata	2006	ERS070815
Prevac4514	Kolkata	2006	ERS070816
Prevac4516	Kolkata	2006	ERS070817
Prevac4518	Kolkata	2006	ERS070818

Prevac4519	Kolkata	2006	ERS070819
Prevac4521	Kolkata	2006	ERS070820
Prevac4523	Kolkata	2006	ERS070821
Prevac4544	Kolkata	2007	ERS070822
Prevac4547	Kolkata	2007	ERS070823
Prevac4624	Kolkata	2007	ERS070824
Prevac4630	Kolkata	2003	ERS070825
Prevac4631	Kolkata	2004	ERS070826
Prevac4632	Kolkata	2004	ERS070827
Prevac4634	Kolkata	2005	ERS070828
Prevac4635	Kolkata	2005	ERS070829
Prevac4636	Kolkata	2005	ERS070830
Prevac4637	Kolkata	2005	ERS070831
Prevac4638	Kolkata	2006	ERS070832
Prevac4639	Kolkata	2006	ERS070833
Prevac4640	Kolkata	2006	ERS070834
Prevac4641	Kolkata	2006	ERS070835
Prevac4642	Kolkata	2006	ERS070836
Prevac4643	Kolkata	2006	ERS070837
Prevac4644	Kolkata	2006	ERS070838
Prevac4648	Kolkata	2003	ERS070839
Prevac4649	Kolkata	2003	ERS070840
82	Mexico	1998	ERS066572
838	Mexico	1999	ERS066573
54a	Mexico	1999	ERS066574
85	Mexico	2000	ERS066578
2006	Mexico	2004	ERS066591
5032	Mexico	2005	ERS066592
688	Mexico	2006	ERS066593
353	Mexico	2008	ERS066595
EM-0892	Mexico	2002	ERS066596
Prevac4650	Kolkata	2003	ERS070841
Prevac4651	Kolkata	2004	ERS070842
Prevac4652	Kolkata	2004	ERS070843
Prevac4653	Kolkata	2005	ERS070844
Prevac4654	Kolkata	2005	ERS070845
Prevac4655	Kolkata	2006	ERS070846
Prevac4656	Kolkata	2006	ERS070847
Prevac4657	Kolkata	2006	ERS070848
Prevac4236	Kolkata	2003	ERS070553
Prevac4237	Kolkata	2003	ERS070554
Prevac4238	Kolkata	2003	ERS070555
Prevac4239	Kolkata	2003	ERS070556
Prevac4240	Kolkata	2003	ERS070557

Prevac4241	Kolkata	2003	ERS070558
Prevac4243	Kolkata	2003	ERS070560
Prevac4244	Kolkata	2003	ERS070561
Prevac4245	Kolkata	2003	ERS070562
Prevac4246	Kolkata	2003	ERS070563
Prevac4247	Kolkata	2003	ERS070564
Prevac4248	Kolkata	2003	ERS070565
Prevac4249	Kolkata	2003	ERS070566
Prevac4250	Kolkata	2003	ERS070567
Prevac4251	Kolkata	2003	ERS070568
Prevac4252	Kolkata	2003	ERS070569
Prevac4253	Kolkata	2003	ERS070570
Prevac4254	Kolkata	2003	ERS070571
Prevac4255	Kolkata	2003	ERS070572
Prevac4256	Kolkata	2003	ERS070573
Prevac4257	Kolkata	2003	ERS070574
Prevac4258	Kolkata	2003	ERS070575
Prevac4259	Kolkata	2003	ERS070576
Prevac4260	Kolkata	2003	ERS070577
Prevac4261	Kolkata	2003	ERS070578
Prevac4262	Kolkata	2003	ERS070579
Prevac4263	Kolkata	2003	ERS070580
Prevac4264	Kolkata	2003	ERS070581
Prevac4265	Kolkata	2003	ERS070582
Prevac4267	Kolkata	2003	ERS070584
Prevac4268	Kolkata	2003	ERS070585
Prevac4269	Kolkata	2003	ERS070586
Prevac4270	Kolkata	2003	ERS070587
Prevac4271	Kolkata	2003	ERS070588
Prevac4272	Kolkata	2003	ERS070589
Prevac4273	Kolkata	2003	ERS070590
Prevac4274	Kolkata	2003	ERS070591
Prevac4275	Kolkata	2003	ERS070592
Prevac4276	Kolkata	2003	ERS070593
Prevac4277	Kolkata	2003	ERS070594
Prevac4278	Kolkata	2003	ERS070595
Prevac4279	Kolkata	2003	ERS070596
Prevac4281	Kolkata	2003	ERS070597
Prevac4282	Kolkata	2003	ERS070598
Prevac4283	Kolkata	2003	ERS070599
Prevac4284	Kolkata	2003	ERS070600
Prevac4285	Kolkata	2003	ERS070601
Prevac4286	Kolkata	2003	ERS070602
Prevac4287	Kolkata	2003	ERS070603

Prevac4288	Kolkata	2003	ERS070604
Prevac4289	Kolkata	2003	ERS070605
Prevac4290	Kolkata	2003	ERS070606
Prevac4291	Kolkata	2003	ERS070607
Prevac4292	Kolkata	2003	ERS070608
Prevac4293	Kolkata	2003	ERS070609
Prevac4294	Kolkata	2003	ERS070610
Prevac4296	Kolkata	2003	ERS070611
Prevac4297	Kolkata	2003	ERS070612
Prevac4299	Kolkata	2003	ERS070613
Prevac4300	Kolkata	2003	ERS070614
Prevac4302	Kolkata	2003	ERS070615
Prevac4303	Kolkata	2003	ERS070616
Prevac4304	Kolkata	2004	ERS070617
Prevac4305	Kolkata	2004	ERS070618
Prevac4306	Kolkata	2004	ERS070619
Prevac4307	Kolkata	2004	ERS070620
Prevac4308	Kolkata	2004	ERS070621
Prevac4309	Kolkata	2004	ERS070622
Prevac4310	Kolkata	2004	ERS070623
Prevac4311	Kolkata	2004	ERS070624
Prevac4312	Kolkata	2004	ERS070625
Prevac4313	Kolkata	2004	ERS070626
Prevac4314	Kolkata	2004	ERS070627
Prevac4315	Kolkata	2004	ERS070628
Prevac4316	Kolkata	2004	ERS070629
Prevac4318	Kolkata	2004	ERS070630
Prevac4319	Kolkata	2004	ERS070631
Prevac4320	Kolkata	2004	ERS070632
Prevac4322	Kolkata	2004	ERS070633
Prevac4323	Kolkata	2004	ERS070634
Prevac4324	Kolkata	2004	ERS070635
Prevac4325	Kolkata	2004	ERS070636
Prevac4326	Kolkata	2004	ERS070637
Prevac4327	Kolkata	2004	ERS070638
Prevac4328	Kolkata	2004	ERS070639
Prevac4329	Kolkata	2004	ERS070640
Prevac4330	Kolkata	2004	ERS070641
Prevac4331	Kolkata	2004	ERS070642
Prevac4332	Kolkata	2004	ERS070643
Prevac4333	Kolkata	2004	ERS070644
Prevac4334	Kolkata	2004	ERS070645
Prevac4335	Kolkata	2004	ERS070646
Prevac4336	Kolkata	2004	ERS070647

Prevac4337	Kolkata	2004	ERS070648
VE1	Kenya	2009	ERS025600
VE2	Kenya	2009	ERS025601
VE3	Kenya	2009	ERS025602
KNE3C	Kenya	2010	ERS075006
KNE134	Kenya	2009	ERS075013
KNE134B	Kenya	2009	ERS075015
KNE11B	Kenya	2010	ERS075017
KNE170	Kenya	2009	ERS075018
KNE3G	Kenya	2010	ERS075023
KNEXX	Kenya	2009	ERS075077
KNEXC	Kenya	2009	ERS075088
KNEXXH	Kenya	2009	ERS075089
KNE168	Kenya	2009	ERS075102
KNC145	Kenya	2010	ERS075104
KNC135	Kenya	2009	ERS075105
KNC151	Kenya	2010	ERS075106
KNC8884	Kenya	2010	ERS075107
KNC56	Kenya	2010	ERS075110
KNC8880	Kenya	2010	ERS075111
KNC133	Kenya	2007	ERS075112
KNC64	Kenya	2010	ERS075113
KNC8889	Kenya	2010	ERS075114
KNC149	Kenya	2009	ERS075115
KNC158	Kenya	2010	ERS075116
KNC11	Kenya	2010	ERS075119
KNC144	Kenya	2010	ERS075120
KNC147	Kenya	2010	ERS075121
KNC161	Kenya	2010	ERS075123
KNC146	Kenya	2010	ERS075124
KNC157	Kenya	2009	ERS075125
KNC1888	Kenya	2007	ERS075126
KNC156	Kenya	2009	ERS075127
KNC8885	Kenya	2010	ERS075128
KNC124	Kenya	2009	ERS075129
KNC155	Kenya	2010	ERS075130
KNC143	Kenya	2010	ERS075131
KNC8669	Kenya	2010	ERS075132
KNC231	Kenya	2009	ERS075133
KNC205	Kenya	2009	ERS075134
KNC1583	Kenya	2009	ERS075135
KNC241	Kenya	2009	ERS075136
KNC206	Kenya	2009	ERS075137
KNC233	Kenya	2009	ERS075138

KNC8679	Kenya	2008	ERS075139
KNC1420	Kenya	2009	ERS075140
KNC207	Kenya	2009	ERS075142
KNC8675	Kenya	2009	ERS075143
KNC8572	Kenya	2009	ERS075145
KNC1509	Kenya	2009	ERS075146
KNC8673	Kenya	2009	ERS075149
KNC8678	Kenya	2009	ERS075150
KNC208	Kenya	2009	ERS075151
KNC1709	Kenya	2009	ERS075152
SRR308665	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308690	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308691	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308692	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308693	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308703	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308704	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308705	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308706	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308707	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308708	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308709	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308713	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308715	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308716	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308717	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308721	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308722	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308723	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308724	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308725	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308726	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
SRR308727	Nepal	2010	Hendriksen et al 2011, mBIO, 2(4): E00157-11
2009V_1046	Pakistan	2009	Hasan et al 2012, PNAS, 109(29): E2010-7
2009V_1085	India	2009	Hasan et al 2012, PNAS, 109(29): E2010-7
2009V_1096	India	2009	Hasan et al 2012, PNAS, 109(29): E2010-7
2009V_1116	Pakistan	2009	Hasan et al 2012, PNAS, 109(29): E2010-7
2009V_1131	India	2009	Hasan et al 2012, PNAS, 109(29): E2010-7
2010EL_1749	Cameroon	2010	Hasan et al 2012, PNAS, 109(29): E2010-7
2010EL_1961	Haiti	2010	Hasan et al 2012, PNAS, 109(29): E2010-7
2010EL_2010H	Haiti	2010	Hasan et al 2012, PNAS, 109(29): E2010-7
2010EL_2010N	Haiti	2010	Hasan et al 2012, PNAS, 109(29): E2010-7
2010V_1014	Pakistan	2010	Hasan et al 2012, PNAS, 109(29): E2010-7
2011EL_1089	Haiti	2010	Hasan et al 2012, PNAS, 109(29): E2010-7

2011EL_1137	South Africa	2009	Hasan et al 2012, PNAS, 109(29): E2010-7
2011V_1021	DRC	2011	Hasan et al 2012, PNAS, 109(29): E2010-7
3500_05	India	2005	Hasan et al 2012, PNAS, 109(29): E2010-7
3546_06	India	2006	Hasan et al 2012, PNAS, 109(29): E2010-7
3554_08	Nepal	2008	Hasan et al 2012, PNAS, 109(29): E2010-7
3582_05	Pakistan	2005	Hasan et al 2012, PNAS, 109(29): E2010-7
C6706	Peru	1991	Hasan et al 2012, PNAS, 109(29): E2010-7
Indre91_1	Mexico	1991	Hasan et al 2012, PNAS, 109(29): E2010-7
SRR135540	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135543	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135544	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135545	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135546	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135603	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135605	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135620	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR135621	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR190870	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR190877	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191343	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191347	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191349	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191350	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191351	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191363	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191380	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191381	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191383	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191384	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191386	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191389	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191391	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR191719	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR227303	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR227307	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR227309	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR227311	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR227312	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR227324	Thailand	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR227335	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR346409	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR346410	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
SRR346411	Haiti	2010	Reimer et al 2011, EID, 17(11): 2113-21
N16961	Bangladesh	1975	Heidelberg et al 2000, Nature, 406(6795): 477-83

M66_2	Indonesia	1961	Chun et al 2009, PNAS, 106(36): 15442-7