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**	**Location**	**ENA Accession**
	(M/Y)		
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.

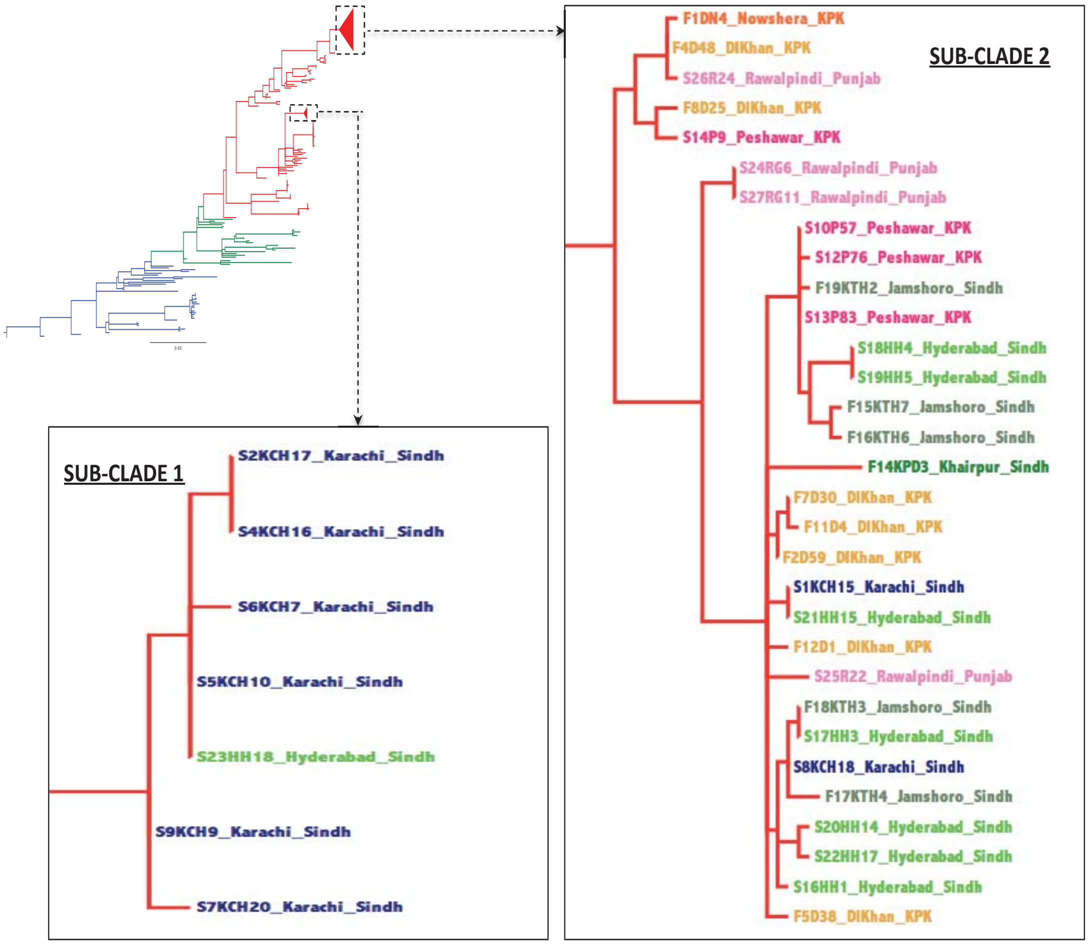


Figure 3.1: Maximum Likelihood phylogenetic tree based on the SNP variation in the Pakistan isolates, showing the relative position of the Pakistani *V. cholerae* O1 El Tor in the wave-3 of the seventh pandemic lineage. The blue, green and red colours of the branches in the tree represent wave-1, 2 and 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).

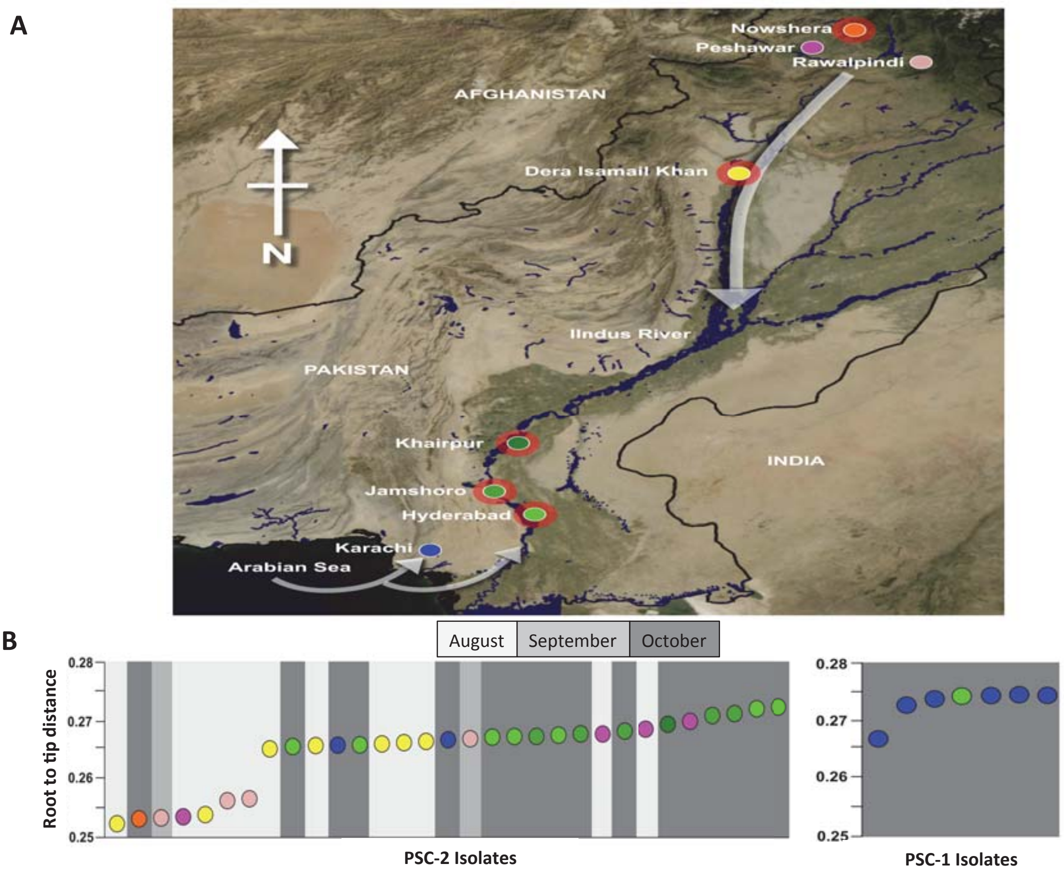


Figure 3.2: A) North orientated map of Pakistan indicating the eight locations where the *V. cholerae* O1 El Tor from this study were isolated (shown by individual coloured circles, red outer shading highlights the five locations that experienced flooding). White arrows show the hypothesised directions 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.

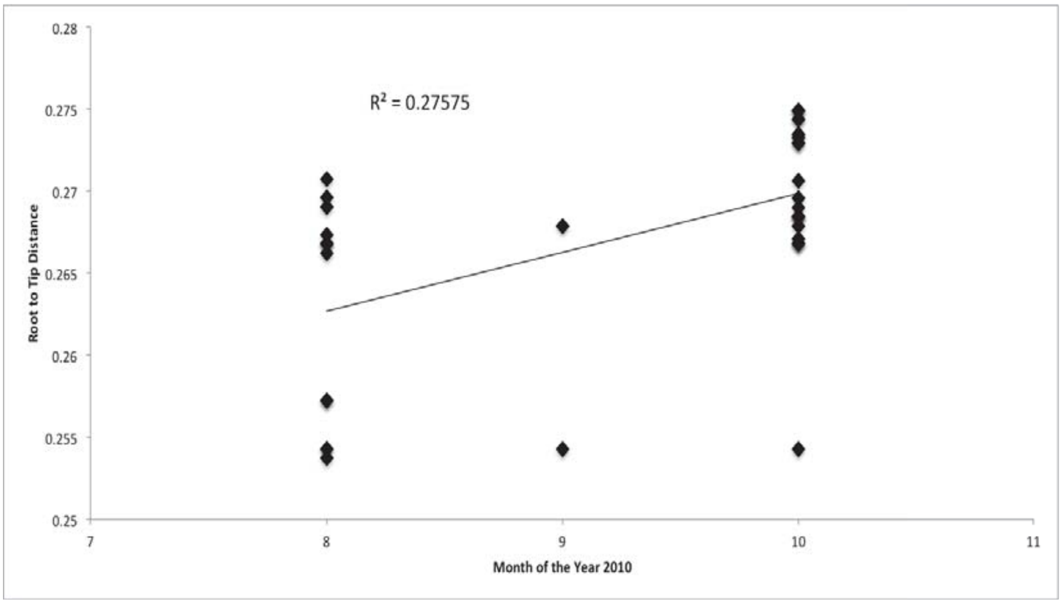


Figure 3.3: A scatter plot of root-to-tip distance *vs.* date of isolation for PSC-1 and PSC-2 combined. The R^2 value represents the correlation between root to tip distance of the Pakistan isolates and their time of isolation.

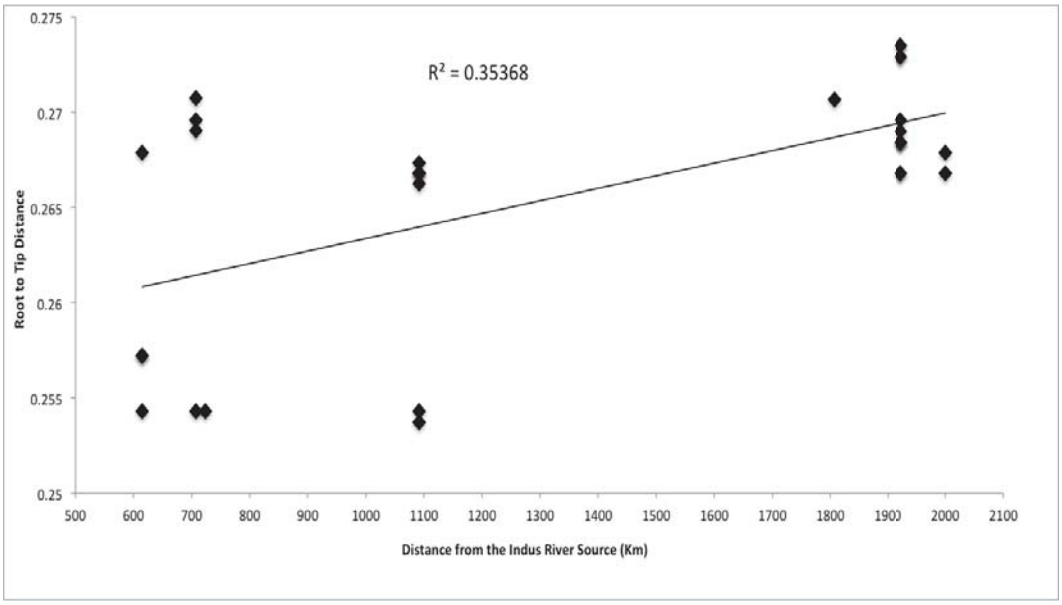


Figure 3.4: A scatter plot of root-to-tip distance *vs.* distance from river source for PSC-2. The R^2 value represents the correlation between root to tip distance of the Pakistan isolates and their distance from the source of the Indus river into Pakistan.

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.

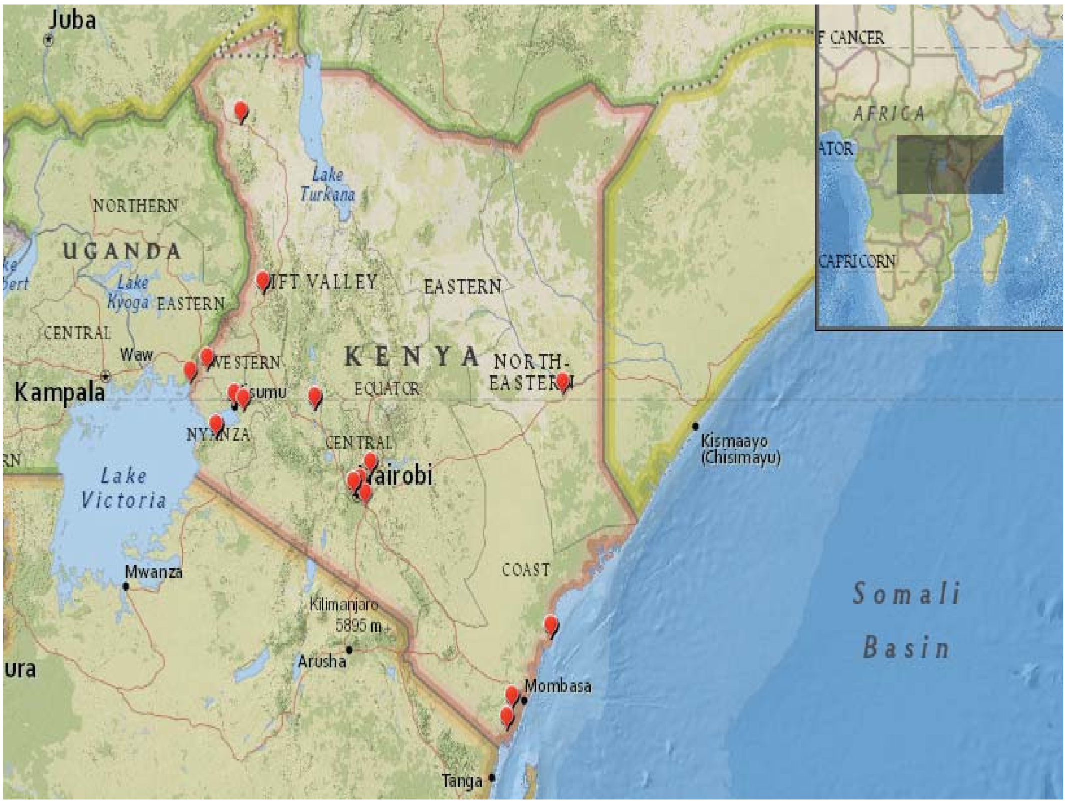


Figure 3.5: The map of Kenya showing the sites from where the isolates of *V. cholerae* were obtained. The inset shows the Kenyan region in context of Africa and each red balloon indicates a sampling site.

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).

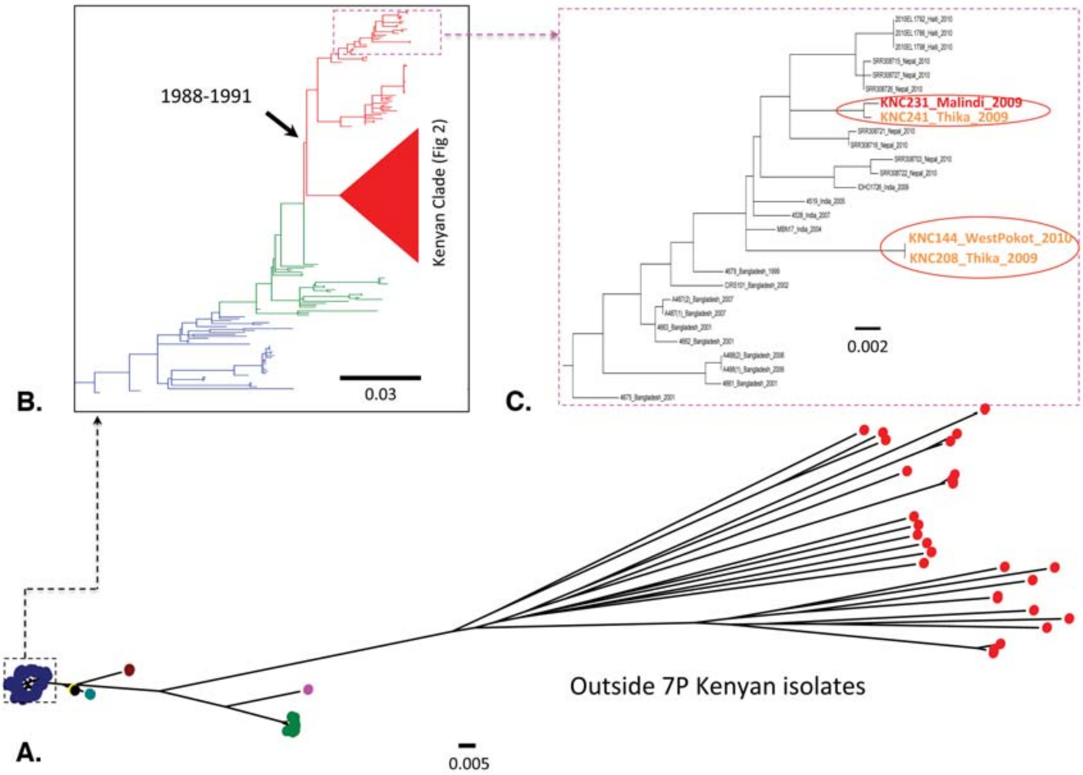


Figure 3.6: A) A Maximum Likelihood phylogenetic tree of *V. cholerae* based on SNP variation. The 6 major O1 clinical groups are shown in this tree with the seventh pandemic El Tor (in blue circles) and classical lineage (in green circles). Environmental non-O1/O139 from Kenya are represented as red 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.

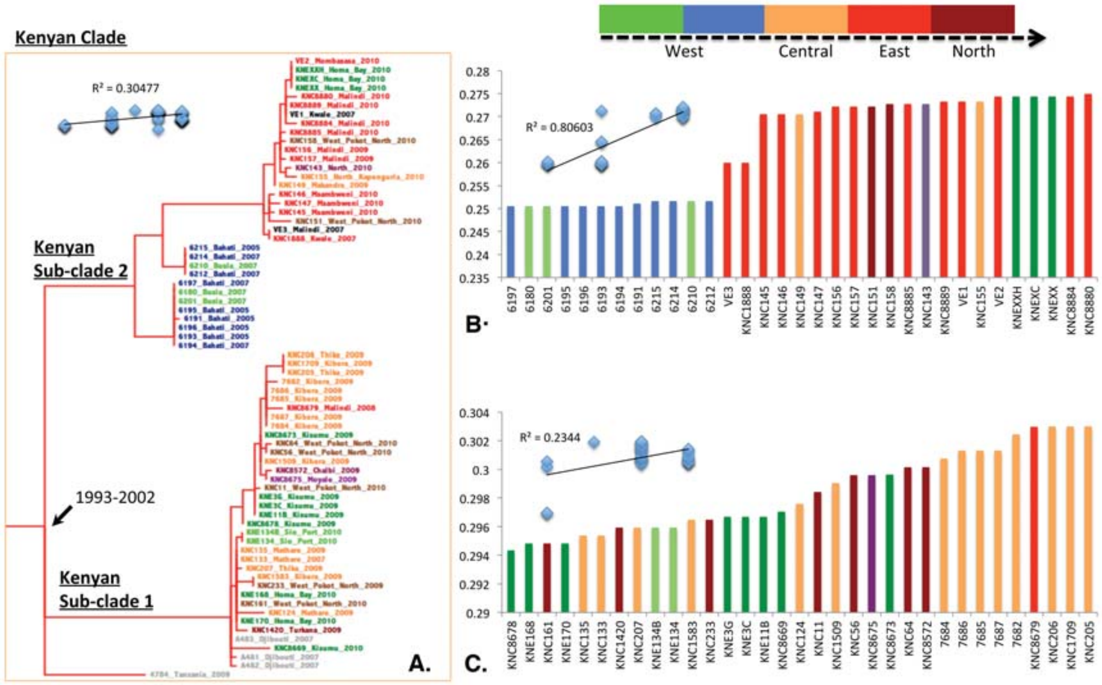


Figure 3.7: A) the exclusive Kenyan clade with two sub-clades, KSC-1 and KSC-2, from within the seventh pandemic maximum likelihood phylogenetic tree. The date on the node represents tMRCA of KSC-1 and 2. B) and C) root to tip distances of isolates of sub-clades KSC-2 and KSC-1 arranged in increasing 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	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE081A	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE102A	🟩	🟩	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE83	🟩	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE59	🟩	🟥	🟥	🟩	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE85	🟥	🟥	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE85C	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE96	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE096B	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE150	🟥	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE10G	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE45	🟥	🟩	🟥	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE70	🟩	🟩	🟥	🟩	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE104C	🟩	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥
KNE7	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE114	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥
KNE60	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥
KNE04C	🟥	🟩	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE109B	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥
KNE109A	🟥	🟥	🟩	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥
KNE53	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥
KNE18	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟩	🟥	🟥
KNE81	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟩
KNE056B_2	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE17	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥
KNE195	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥
KNE083A	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟥	🟩	🟥	🟥	🟥	🟥	🟥	🟥

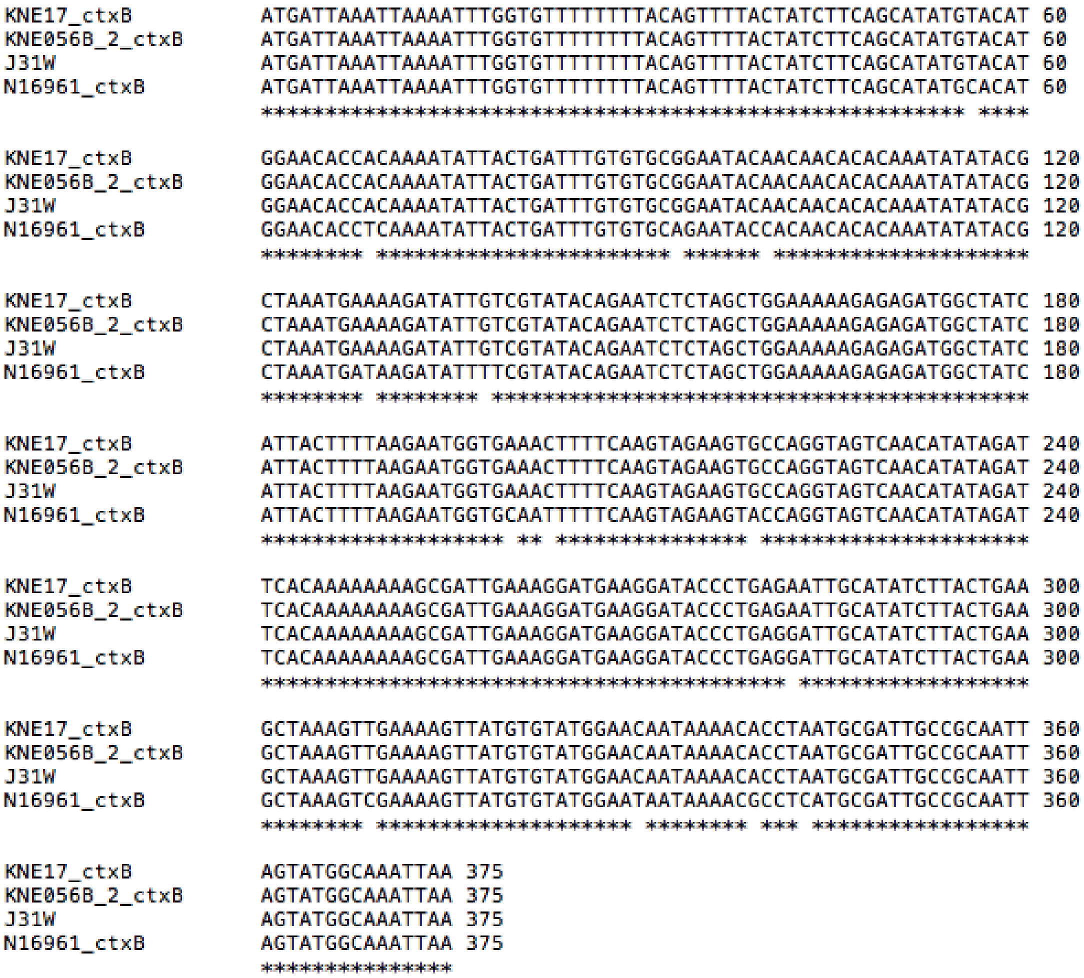
🟥	absent
🟩	present

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          TCACAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGAATTGCATATCTTACTGAA 300
KNE056B_2_ctxB      TCACAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGAATTGCATATCTTACTGAA 300
J31W                TCACAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGCATATCTTACTGAA 300
N16961_ctxB         TCACAAAAAAAAGCGATTGAAAGGATGAAGGATACCCTGAGGATTGCATATCTTACTGAA 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).

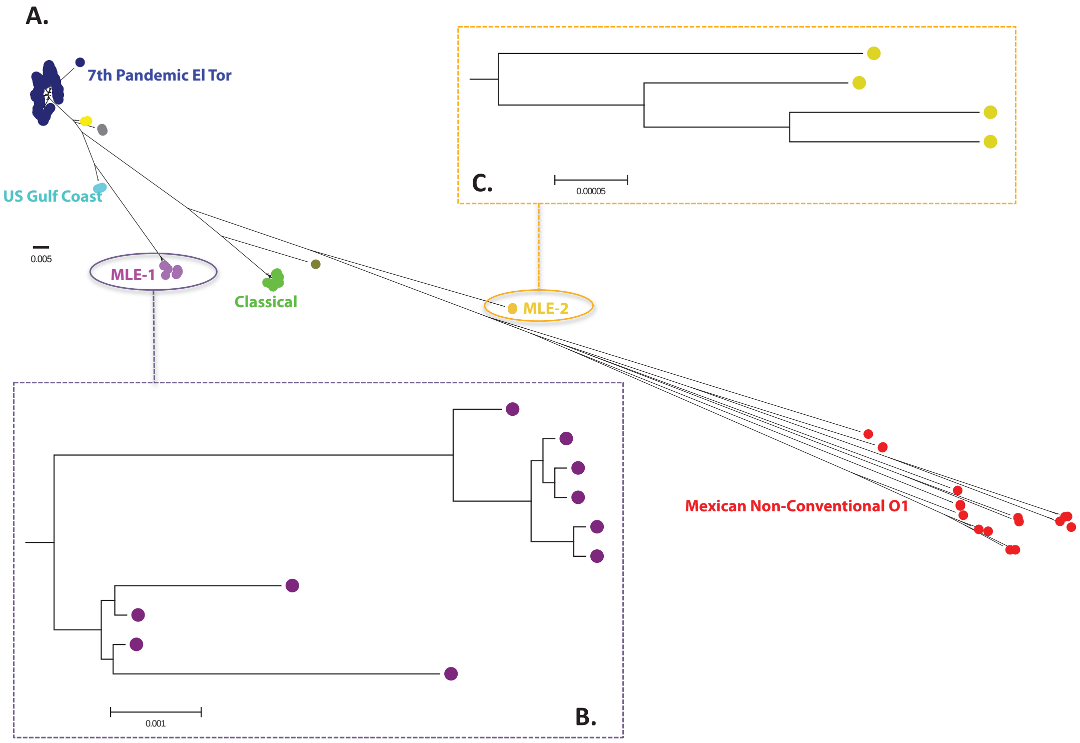


Figure 3.9: A) A Maximum Likelihood phylogenetic tree based on SNP variation. The major lineages are shown and MLE-1 and MLE-2 lineages are circled. In red are the non-conventional O1 isolates from Mexico. B) A mid-point rooted maximum-likelihood phylogenetic tree of the MLE-1 isolates. C) A mid-point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

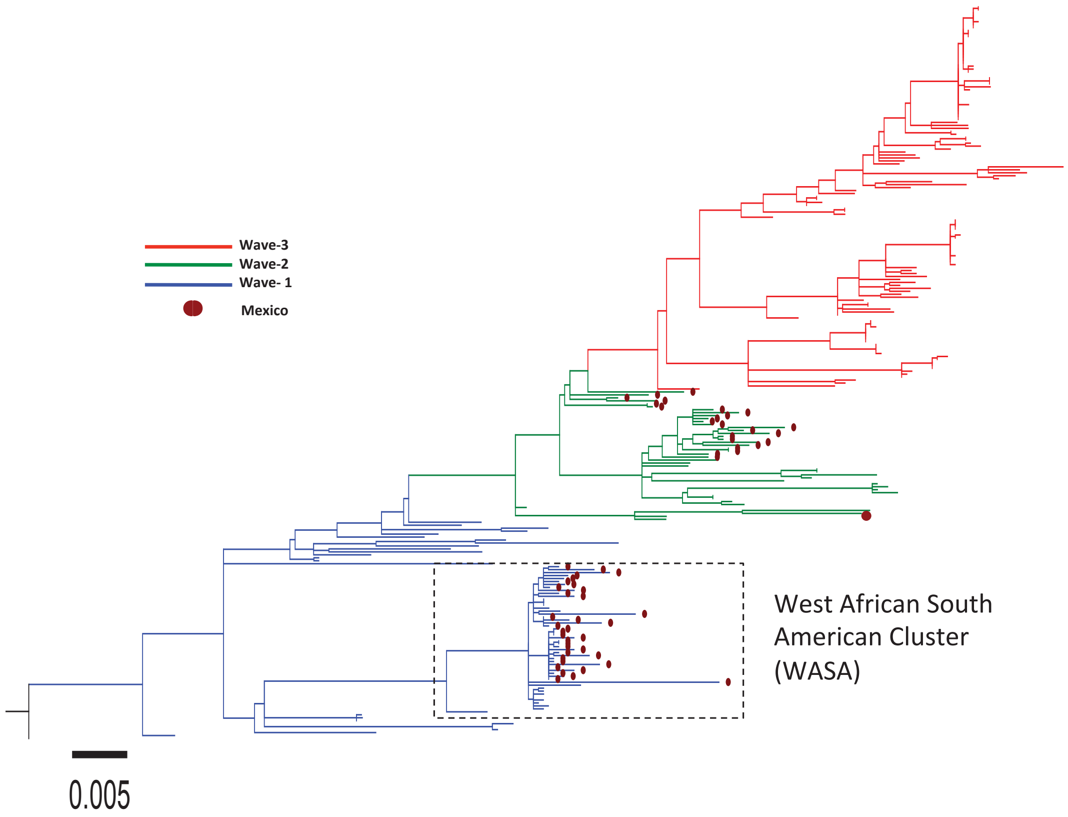


Figure 3.10: Maximum Likelihood phylogenetic tree of the seventh pandemic lineage including the Mexican seventh pandemic isolates from our collection marked as brown circles in the tree. The phylogeny is based on SNP variation after excluding the high density SNPs and likely recombination events. Completed 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.

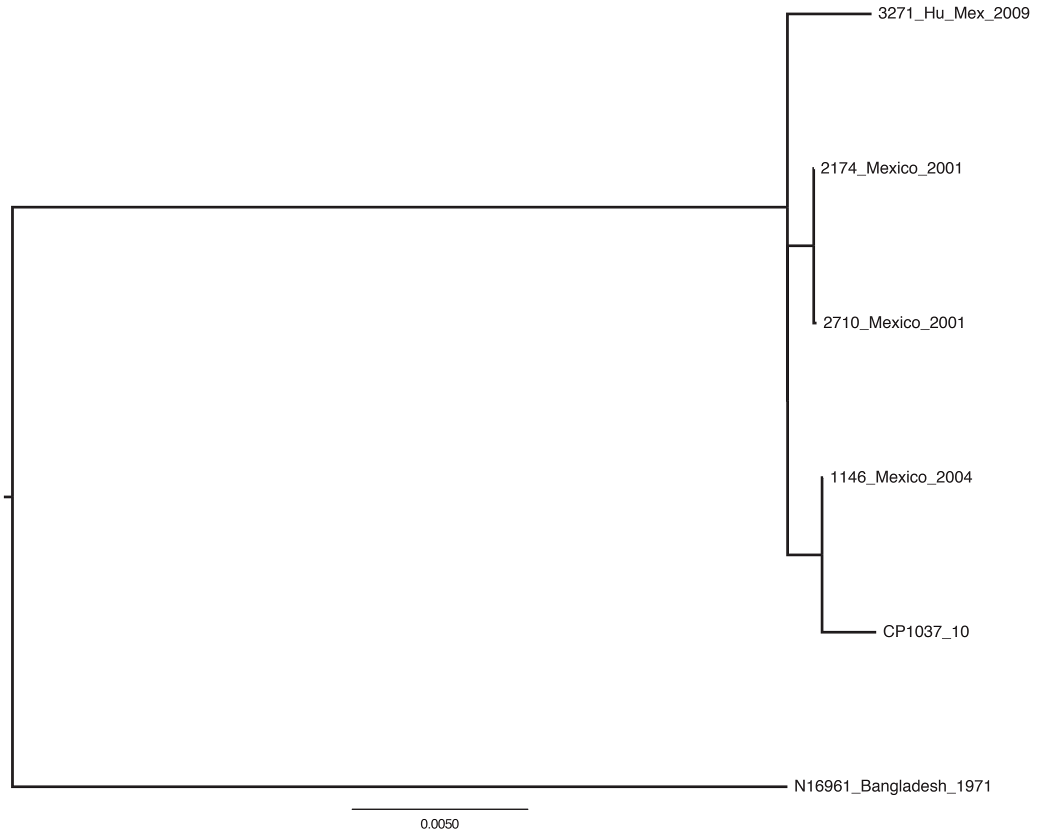


Figure 3.12: Maximum Likelihood phylogeny showing the four MLE-2 isolates and CP1037(10) from NCBI database search results. N16961 was used as the reference to map the sequence data. The scale is given as the number of substitutions per variable site.

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.