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.

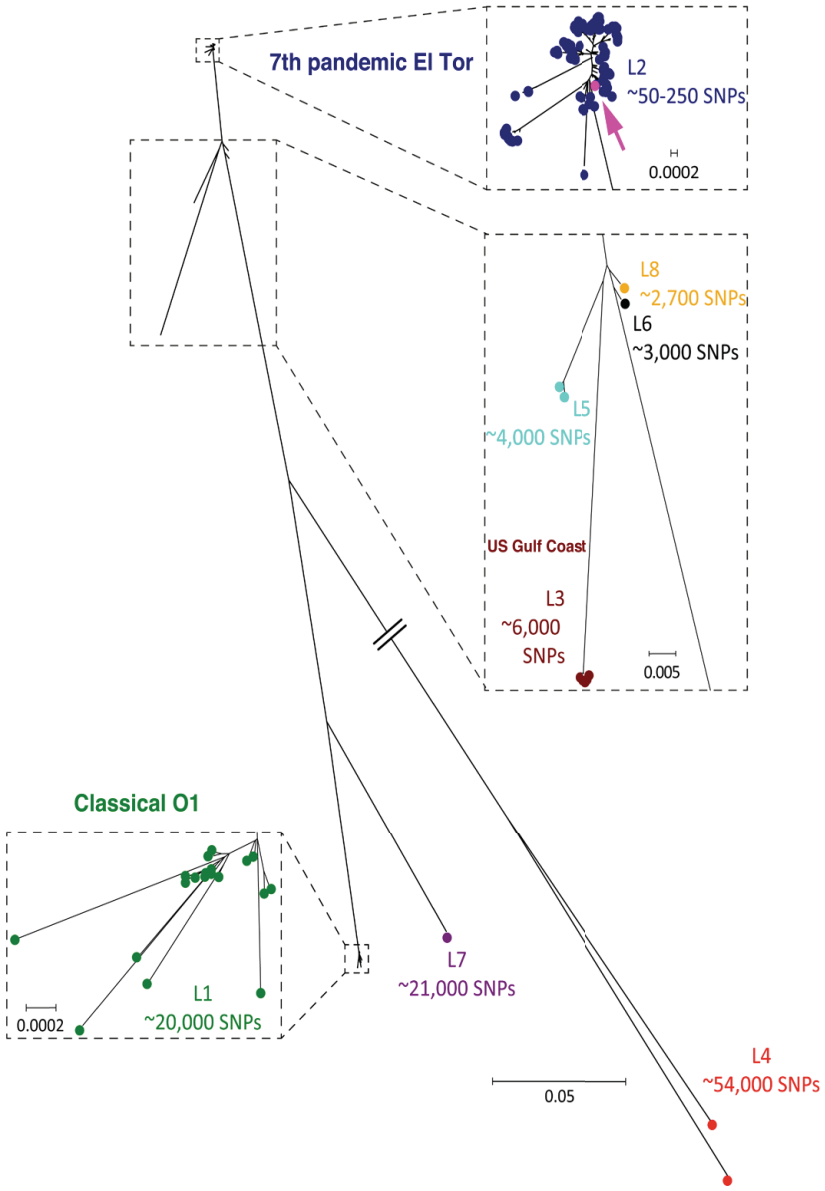


Figure 2.1: Global *V. cholerae* phylogeny of 154 isolates collected between 1910 and 2010. The maximum likelihood tree is based on SNP differences across the whole core genome and the numbers of SNP differences listed are relative to N16961 reference in L2, which is marked with an arrow. The scales 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).

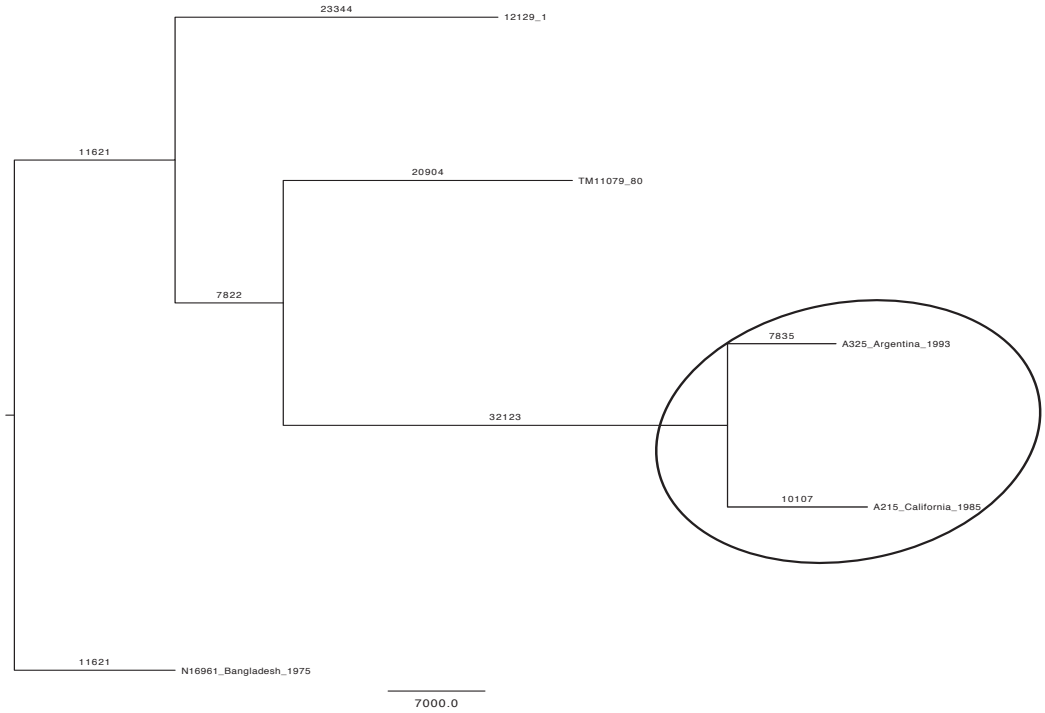


Figure 2.2: Comparison of the "non-conventional O1" strains in our collection with those from Chun *et al.* (see section 2.2). The two strains that are circled are clinical "non-conventional O1" and share common ancestor with TM11079_80. They are separated from the Chun *et al.* environmental strains 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).

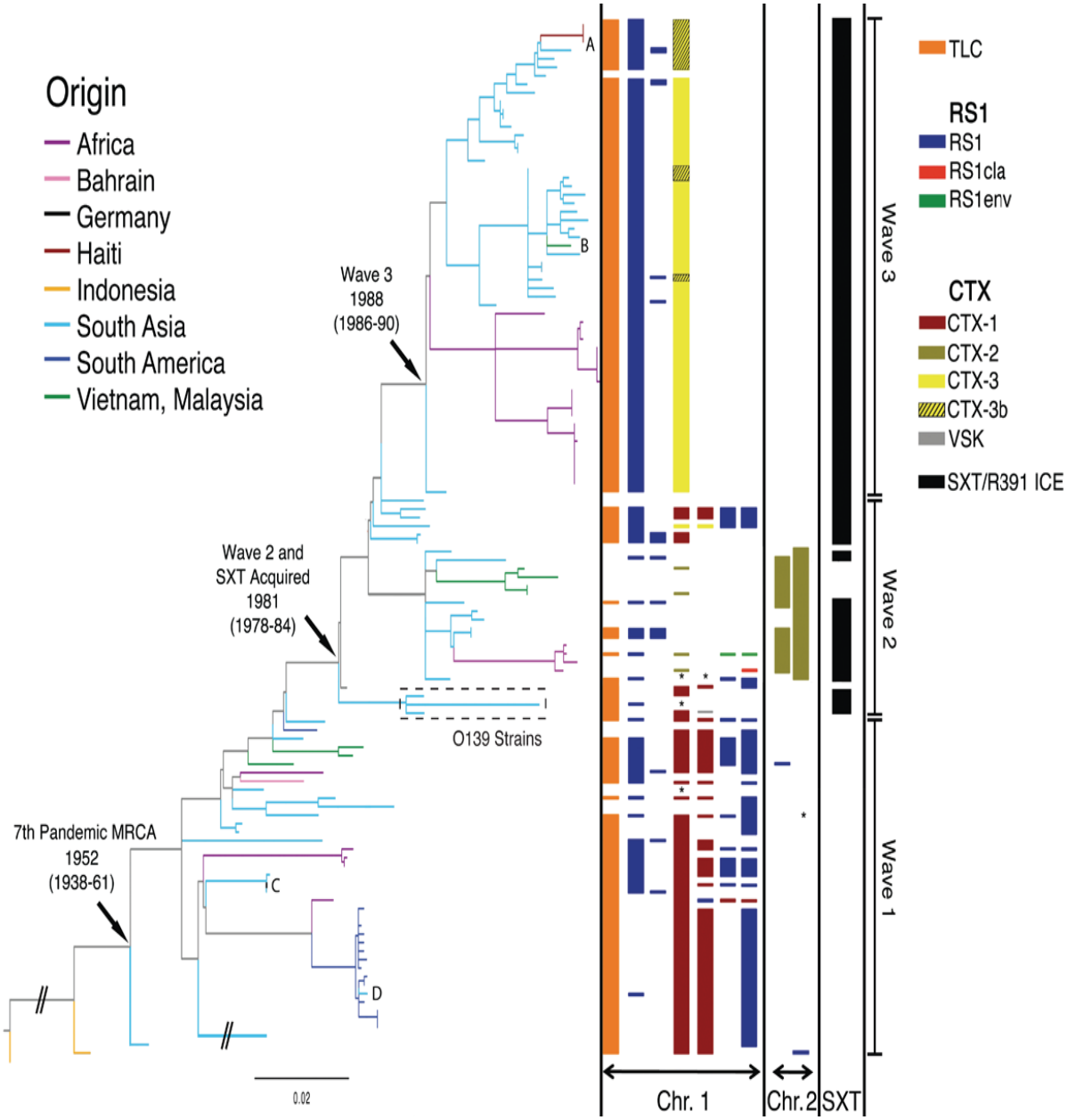


Figure 2.3: Maximum likelihood phylogeny of the seventh pandemic of L2 *V. cholerae* based on SNP differences across the whole genome, excluding likely recombination events. The tree has been rooted using M66 as an out-group and branches are coloured based on the region of isolation of the sample. CTX 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.

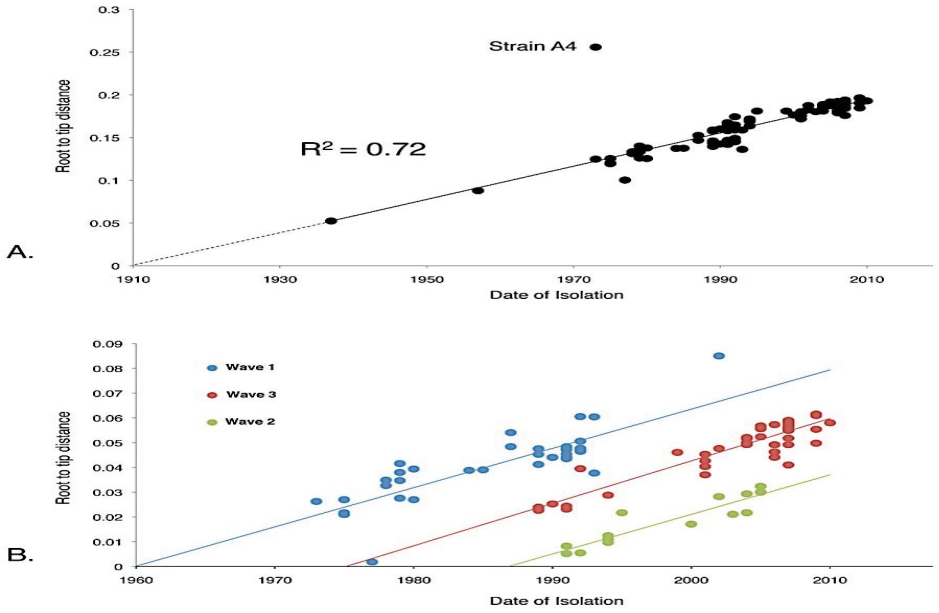


Figure 2.4: A. Root to tip distance of the seventh pandemic strains plotted against time as a linear regression plot **B.** Same analysis plot with each wave plotted separately. Isolate A4, is a 'laboratory strain' that has been multiply passaged that has been removed from the Wave 1 plot.

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).

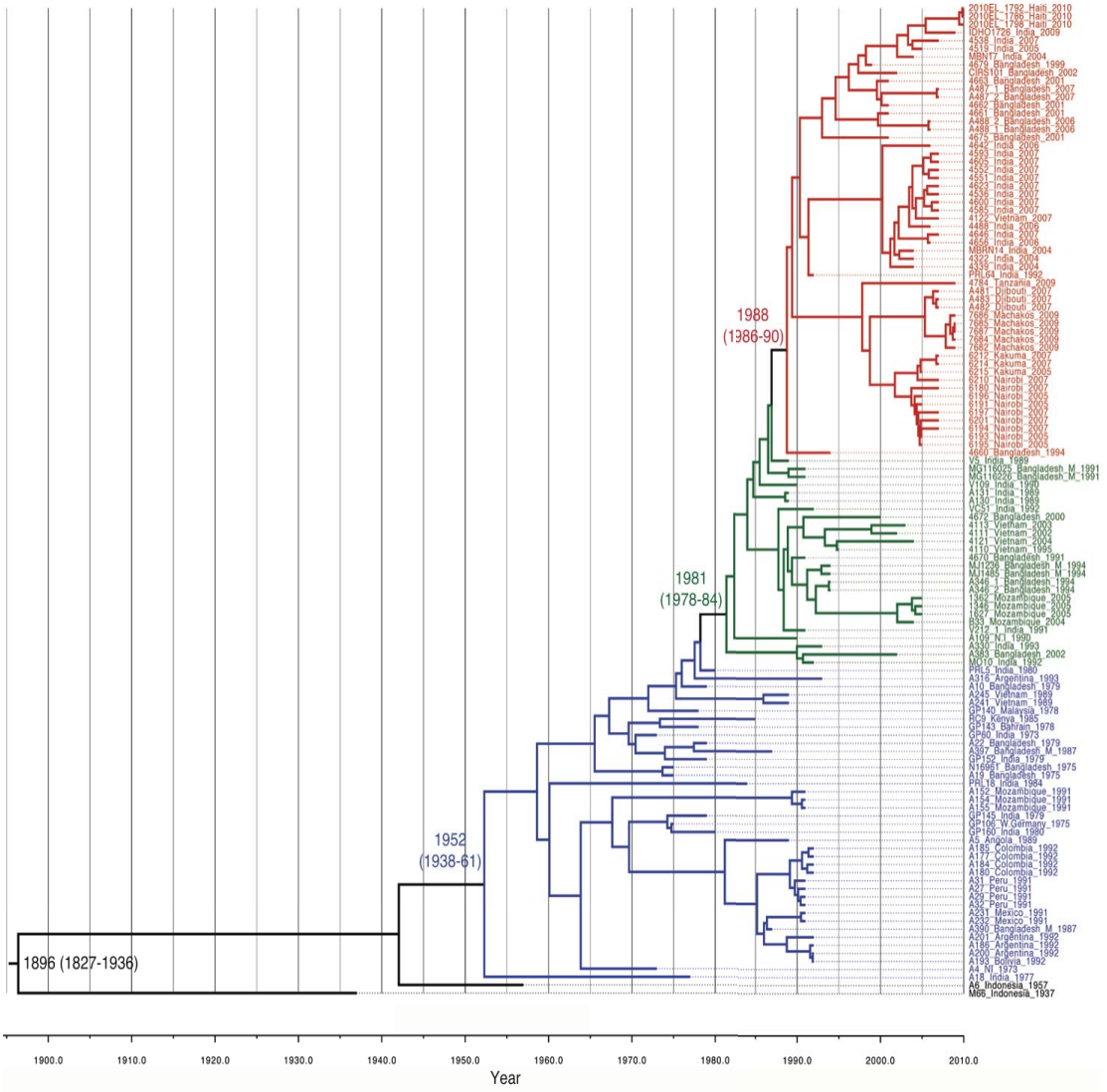


Figure 2.5: BEAST generated maximum clade credibility tree of the seventh cholera pandemic. Dates highlighted are the median values predicted with the date range representing the 95% confidence interval on the estimate. The dates shown are for the MRCAs of the El Tor lineage, wave-1, wave-2/SXT acquisition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

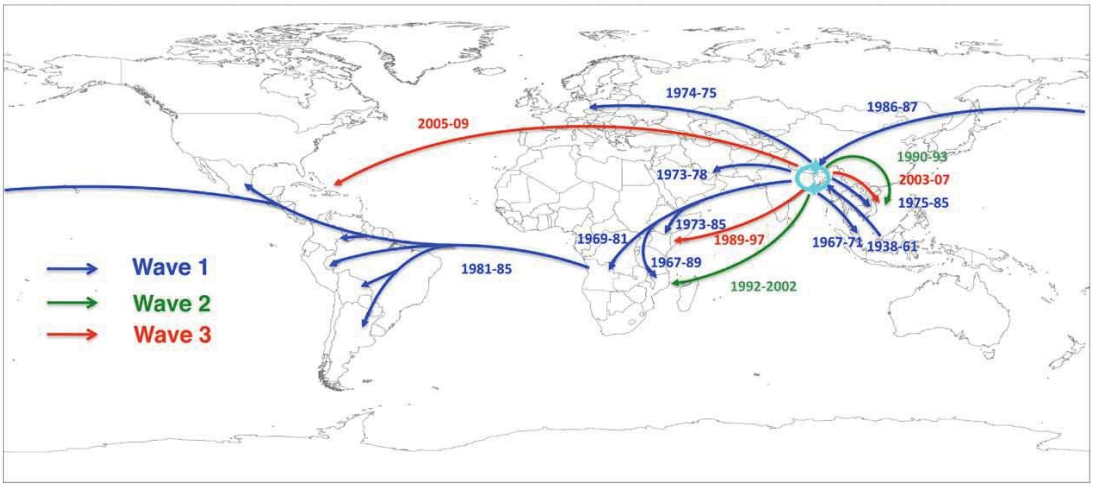


Figure 2.6: seventh Pandemic phylogenetic tree drawn on a global map to show the transmission events. The date ranges shown are derived from Bayesian analysis and represent the median values for the most recent common ancestors of the transmitted strains (later bound), and the MRCA of the transmitted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

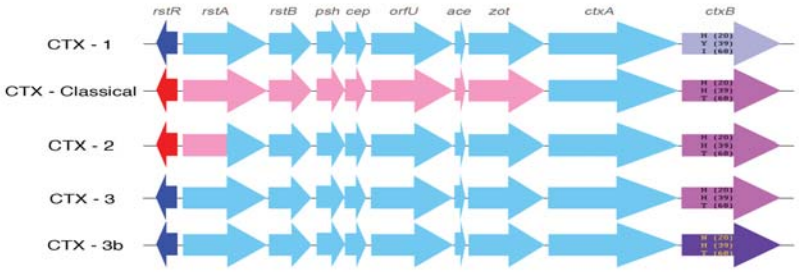
2) For CTX-1 to CTX-3, as there was a shift of *ctx*B$^{\text{El Tor}}$ to *ctx*B$^{\text{Classical}}$, it was called CTX-3.

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	

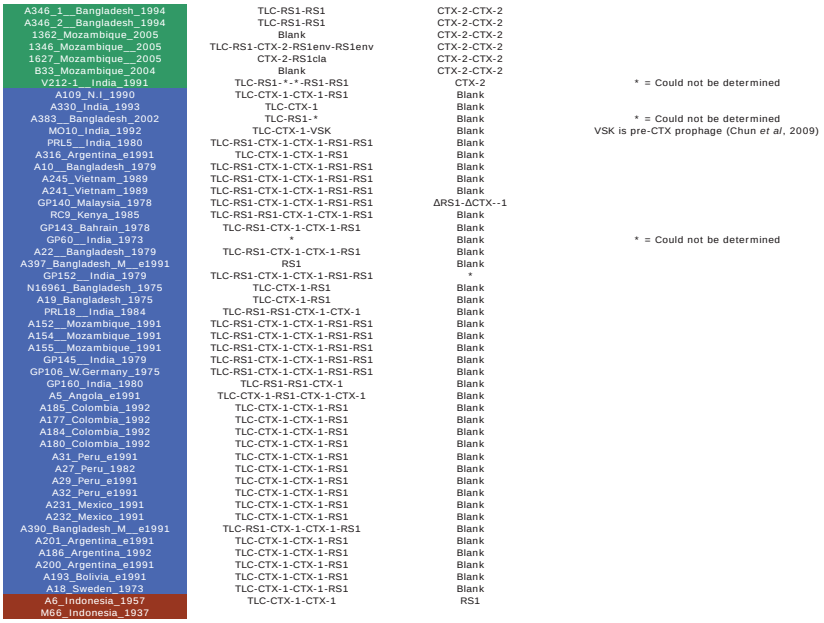


Table 2.2: Structures of CTX types (top panel) and molecular CTX type information of each isolate according to the new nomenclature scheme. In blue, green and red are the wave-1, 2 and 3 seventh pandemic strains with wave and CTX information respectively.

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.

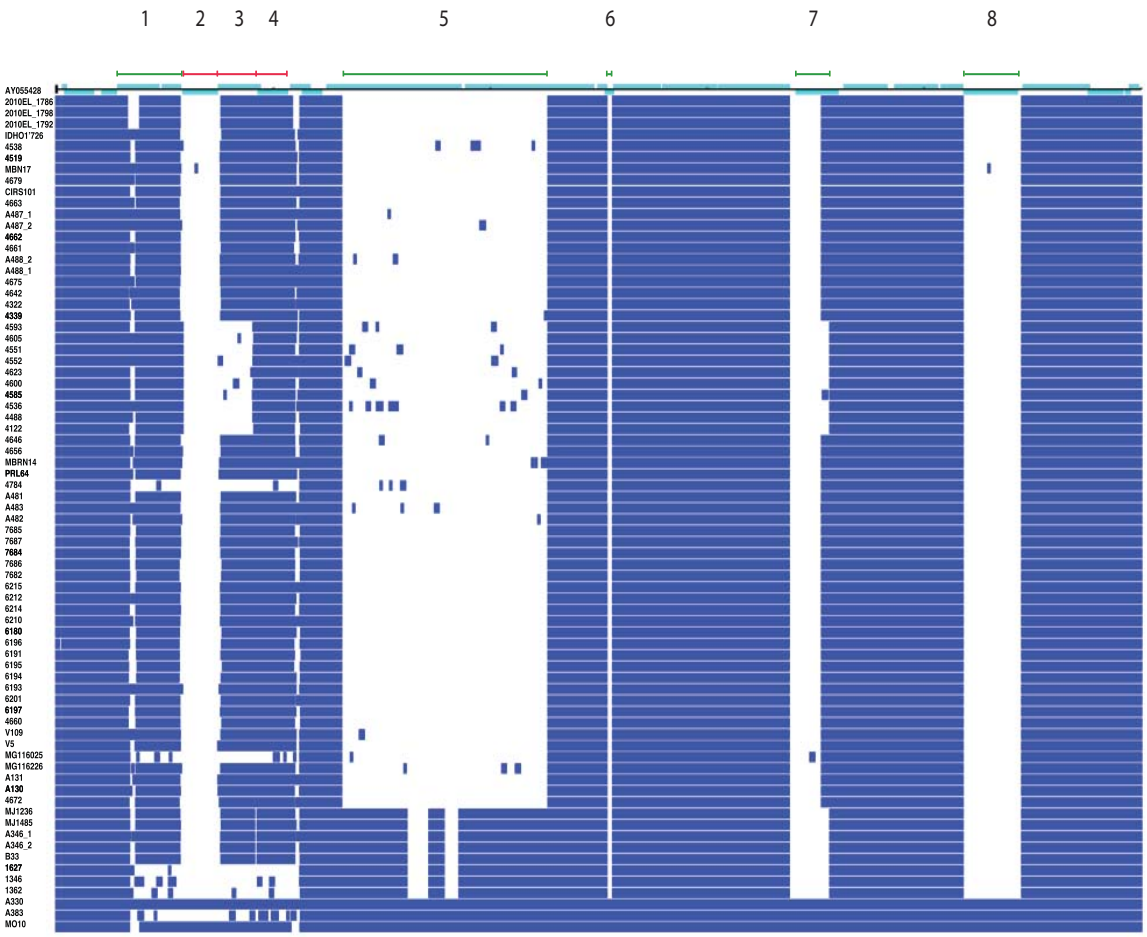


Figure 2.7: Plots showing presence (blue) or absence (white) information of the genes of the SXT of O139 strain MO10 (accession number AY055428) as reference in all the seventh pandemic strains positive for SXT/R391 ICE. The regions marked by green bars are variable and those marked with red bars are 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).

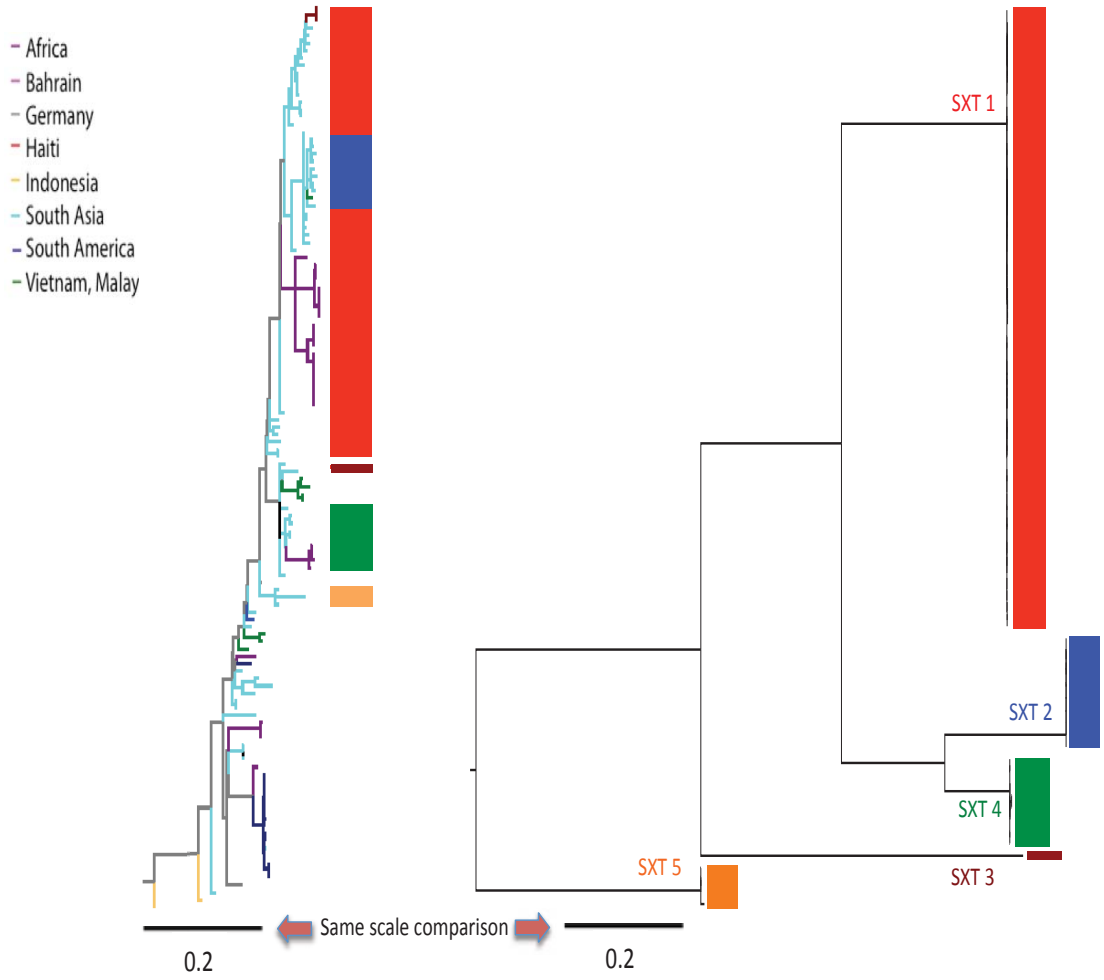


Figure 2.8: Comparison of the maximum likelihood trees of seventh pandemic lineage (left) and core SXT (right). The scales represent substitution per variable site and the colours of the blocks represent the SXT type 1 to 5. Core SXT of MO10, an O139 strain, was used to map the SXT positive isolates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

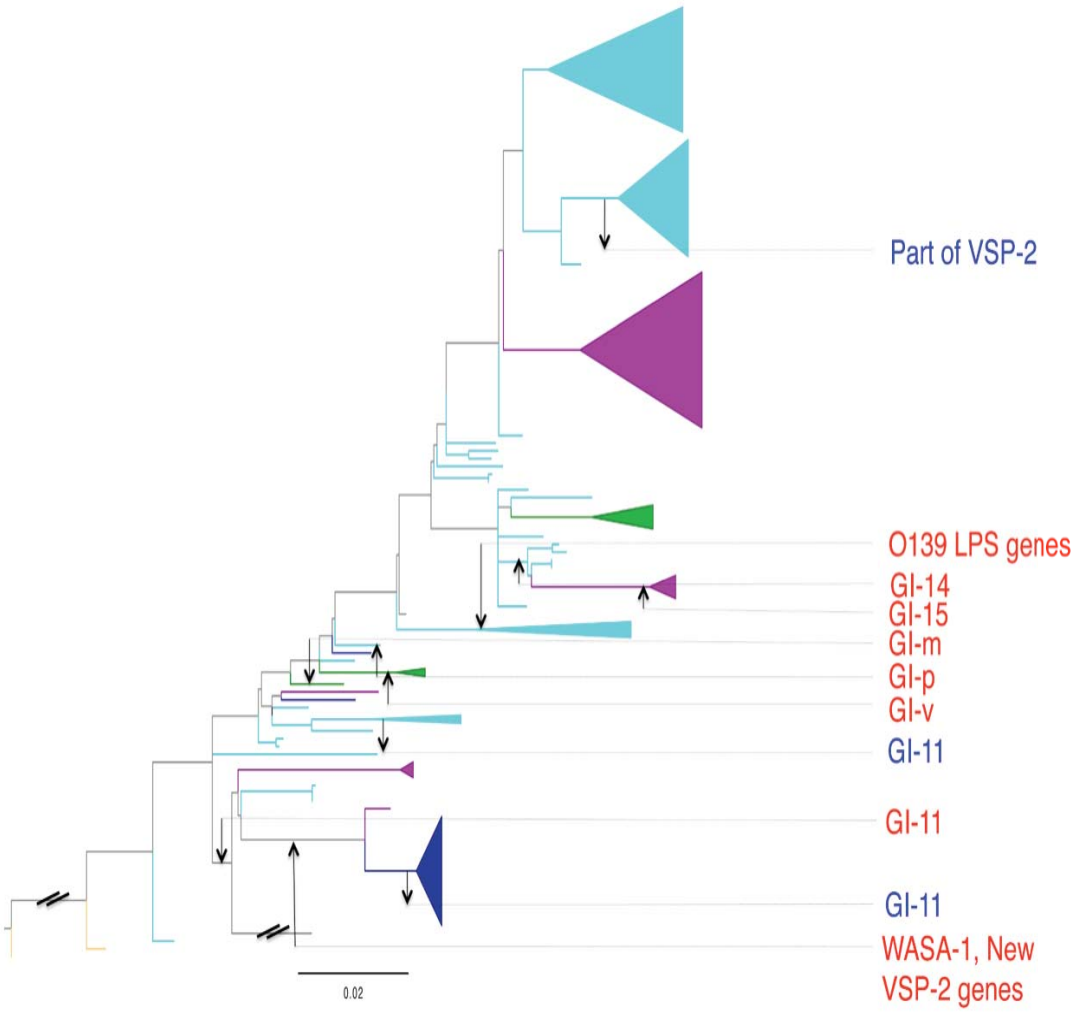


Figure 2.9: Seventh pandemic maximum likelihood tree with the gene flux plotted on the branches. Loci coloured red are insertions and those coloured blue are deletions.

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

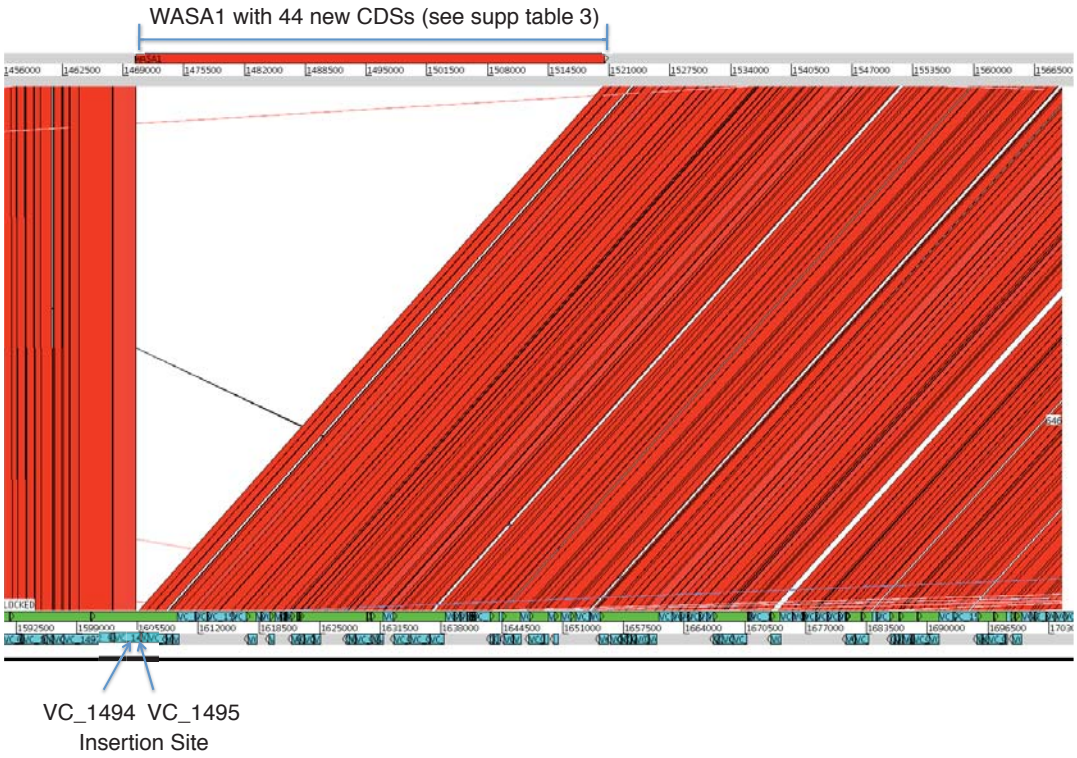


Figure 2.10: Insertion of WASA-1 within VC_1494 (product, aminopeptidase N). The genomes compared are A193 (top) and N16961 (bottom) and the insertion was present in all the WASA strains in our collection.

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).

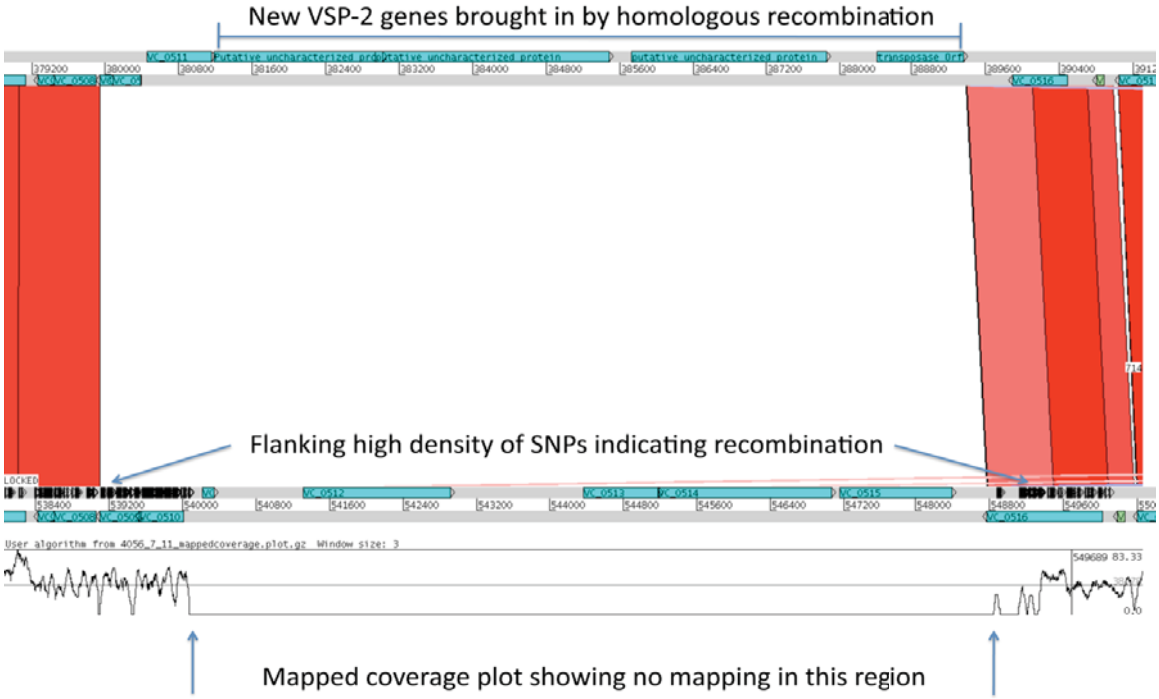


Figure 2.11: Recombination in the WASA cluster with homologous ends showing high SNP density and indicating likely recombination event from outside the tree. The genomes compared are A193 (top) and N16961 (bottom) where A193 represents the WASA cluster.

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.