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

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

5.1 Introduction

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

As *V. cholerae* El Tor isolates are monophyletic, options for the development of classical typing approaches have been relatively limited. One of the approaches of choice was based on differences in sequence within the genes encoding cholera toxin harboured by the CTX phages. However, since CTX are mobile genetic elements, they do not evolve at the same rate or even within the same lineage as the genomic backbone and therefore cannot be trusted for true phylogenetic inference (see section 1.2.11). Molecular approaches, including MLST have little discriminatory power and are also of limited value. Although PCR based approaches based on changes in short repeat elements such as MLVA can detect diversity (Mohamed, *et al.*, 2012) they have limited phylogenetic value. Recently, PFGE typing has become a preferred technique in some national reference centers and public health laboratories. Indeed, it is the current gold standard technique for the sub-typing of *V. cholerae* outbreak isolates. However, again PFGE provides limited phylogenetic information but rather reports on changes in phage elements, restriction sites and broader genome rearrangements. It is also a technique that generates data that is open to interpretation and is difficult to transfer between laboratories.

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

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

The ability to sequence whole bacterial genomes is still limited in many regions of the world where cholera is still occurring. Consequently, methodologies based on simple molecular biology techniques available in routine laboratories that are expandable, relatively economic and reliable were developed to detect these SNPs. Multiplex ligation-dependent probe amplification (MLPA) (MRC-Holland) can be used to detect specific DNA fragments, insertions, deletions, as well as individual SNPs, permiting the detection of multiple targets by amplification driven by a single primer pair in a convenient single reaction.

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

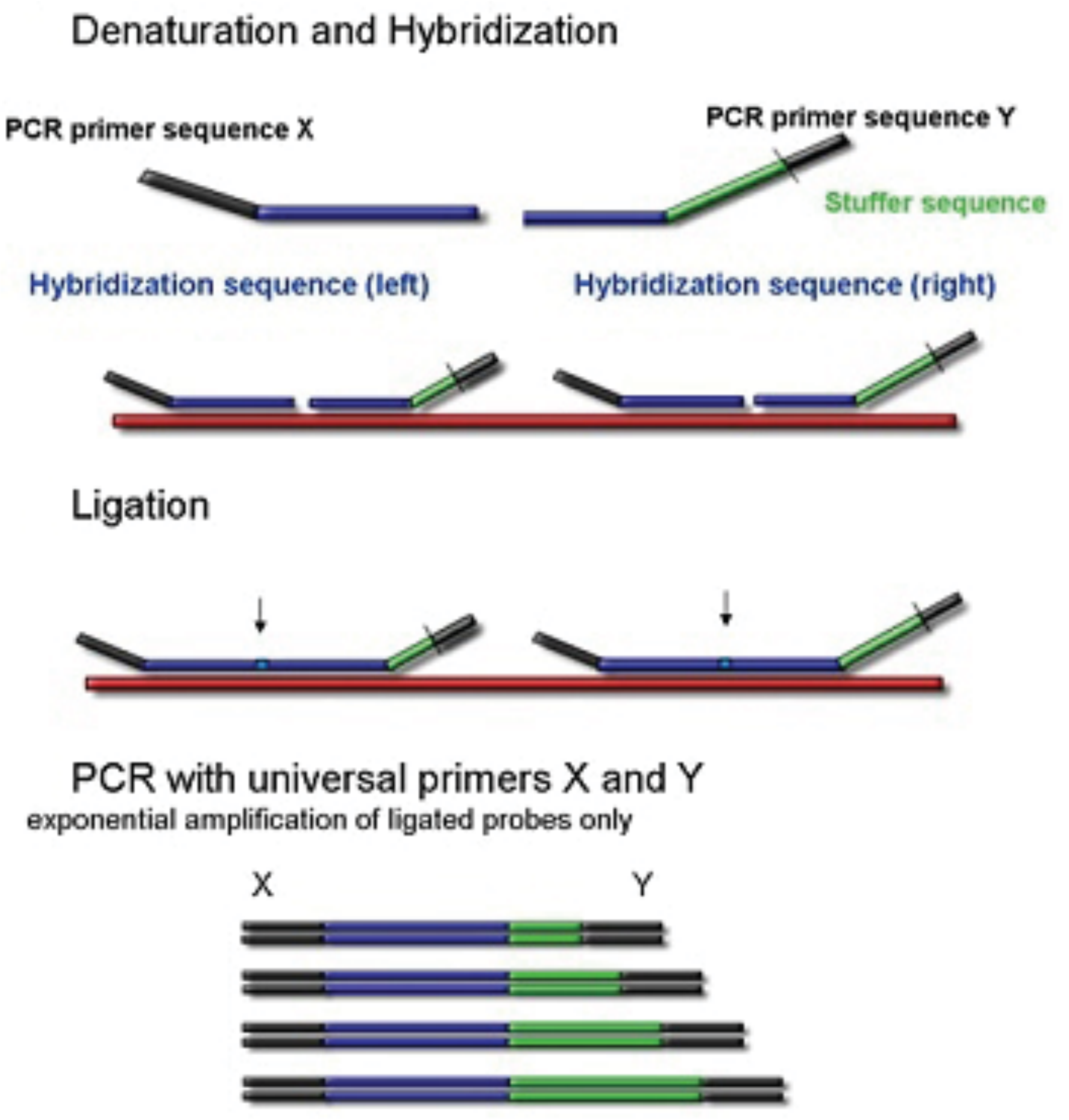

Denaturation and Hybridization
PCR primer sequence X
PCR primer sequence Y
Stuffer sequence
Hybridization sequence (left)
Hybridization sequence (right)
Ligation
PCR with universal primers X and Y
exponential amplification of ligated probes only
X
Y

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

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

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

5.2 Results and discussion

5.2.1 SNPs for genotyping

5.2.1.1 Selection of canonical SNPs

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

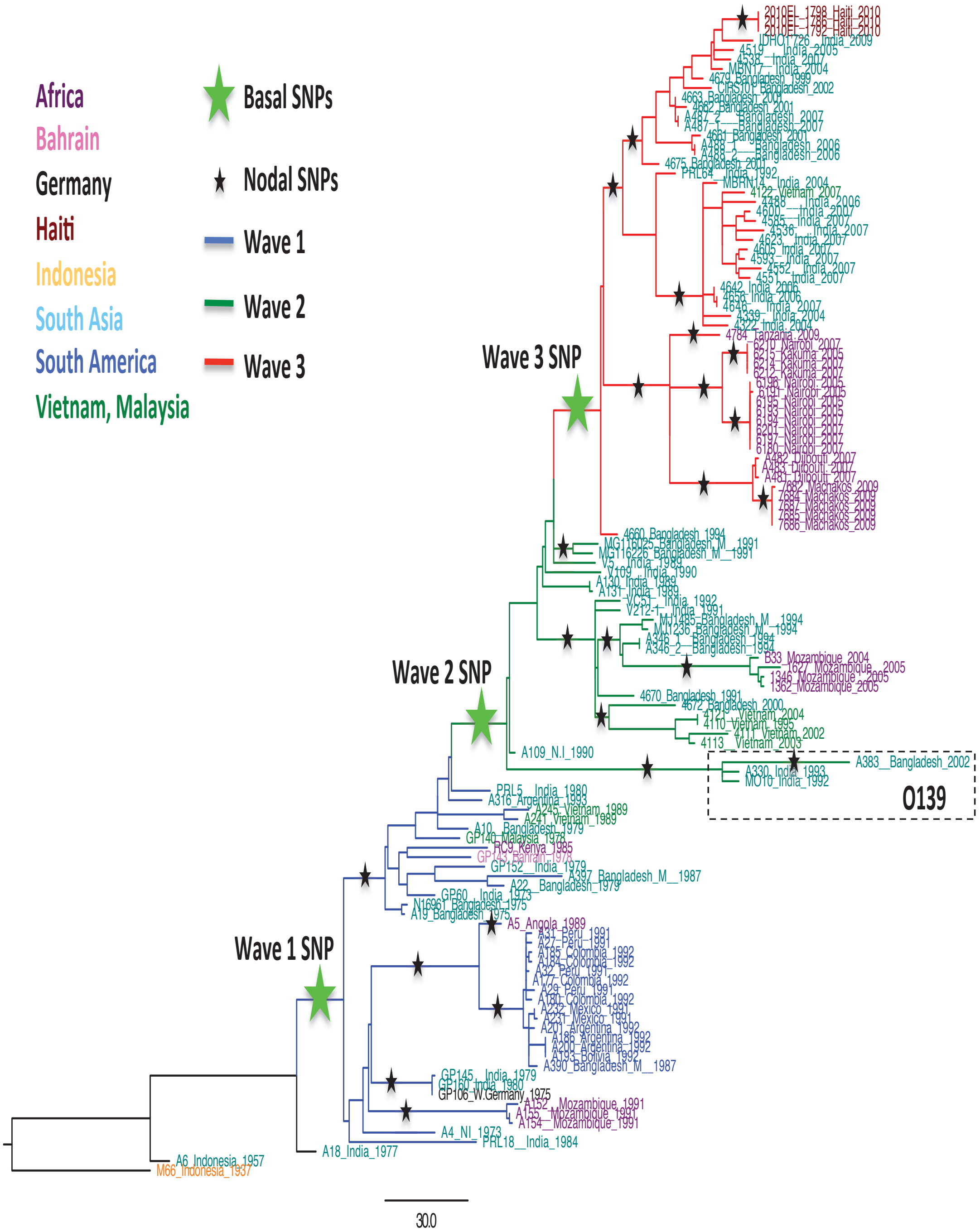


Figure 5.2: A maximum likelihood phylogenetic tree for the global seventh pandemic lineage L2 *V. cholerae* as described in chapter 2. A key to the SNPs selected (represented by stars) for typing is provided in the figure and the scale represents the number of SNPs.

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

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

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

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

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

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

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

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

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

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

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

5.2.1.2 Phylogenetic analysis on selected SNPs

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

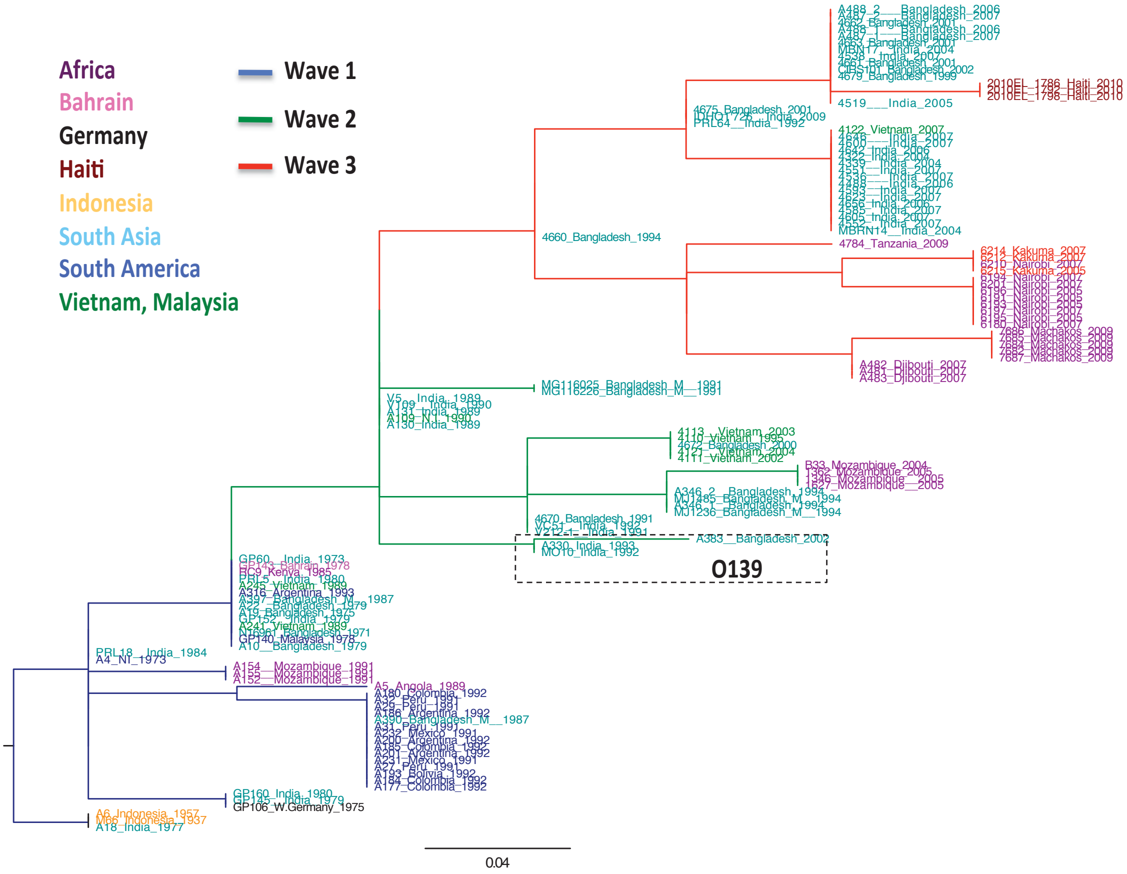


Figure 5.3: A maximum likelihood phylogeny of the 122 *V. cholerae* of L2 lineage described chapter 2, based on the 27 selected SNPs. Each branch is based on a single SNP. A key to locations from which the isolates originated from and the waves to which they belong is provided in the figure and the scale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

5.2.2 Phylogeny expansion and MLPA kits

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

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

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

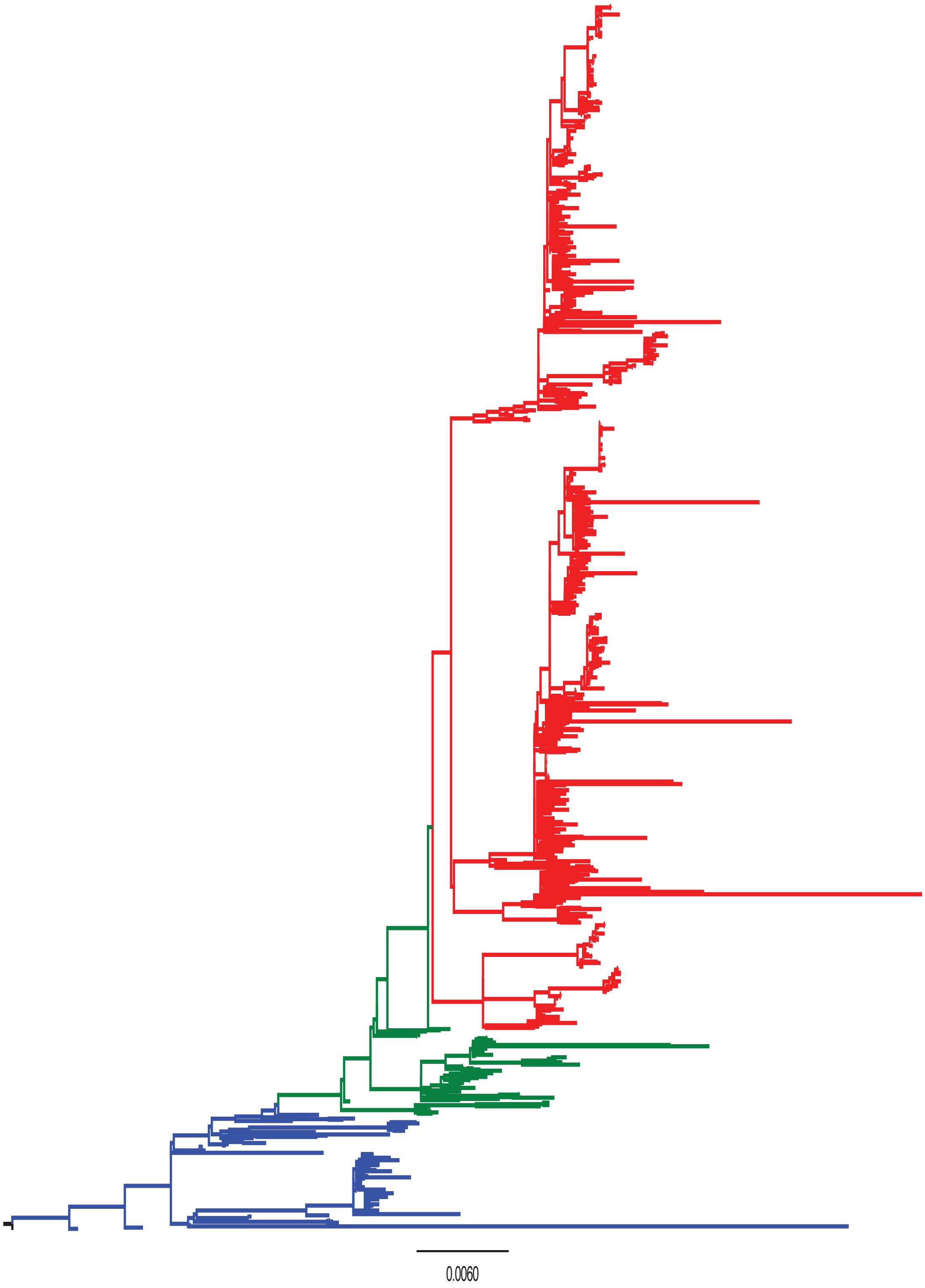


Figure 5.4: The structure of **a** maximum likelihood phylogenetic tree of 1002 seventh pandemic *V. cholerae* El Tor. Blue, green, and red branch colours represent wave-1, 2 and 3 respectively. The reference strain for this tree is *V. cholerae* N16961 El Tor and is rooted using a pre-seventh pandemic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

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

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

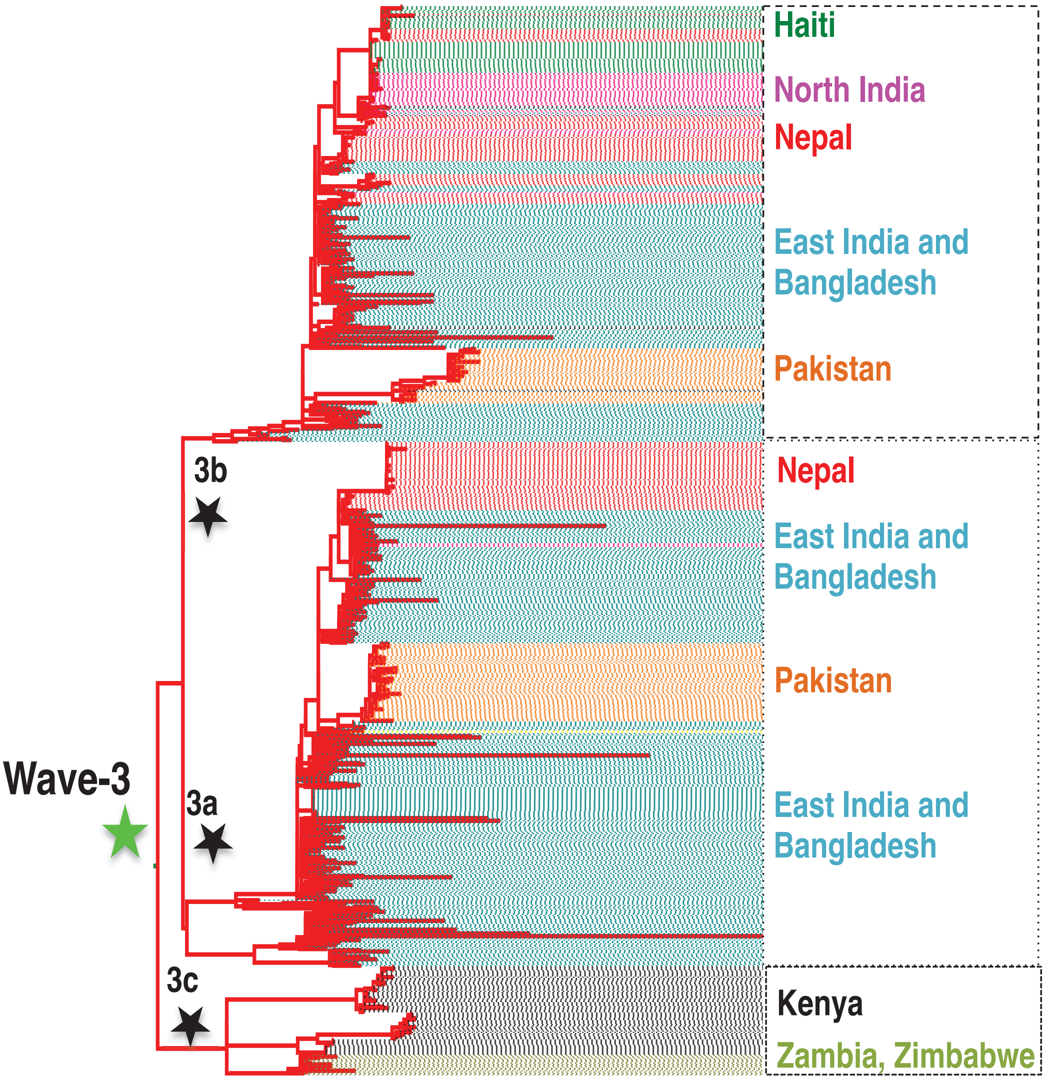
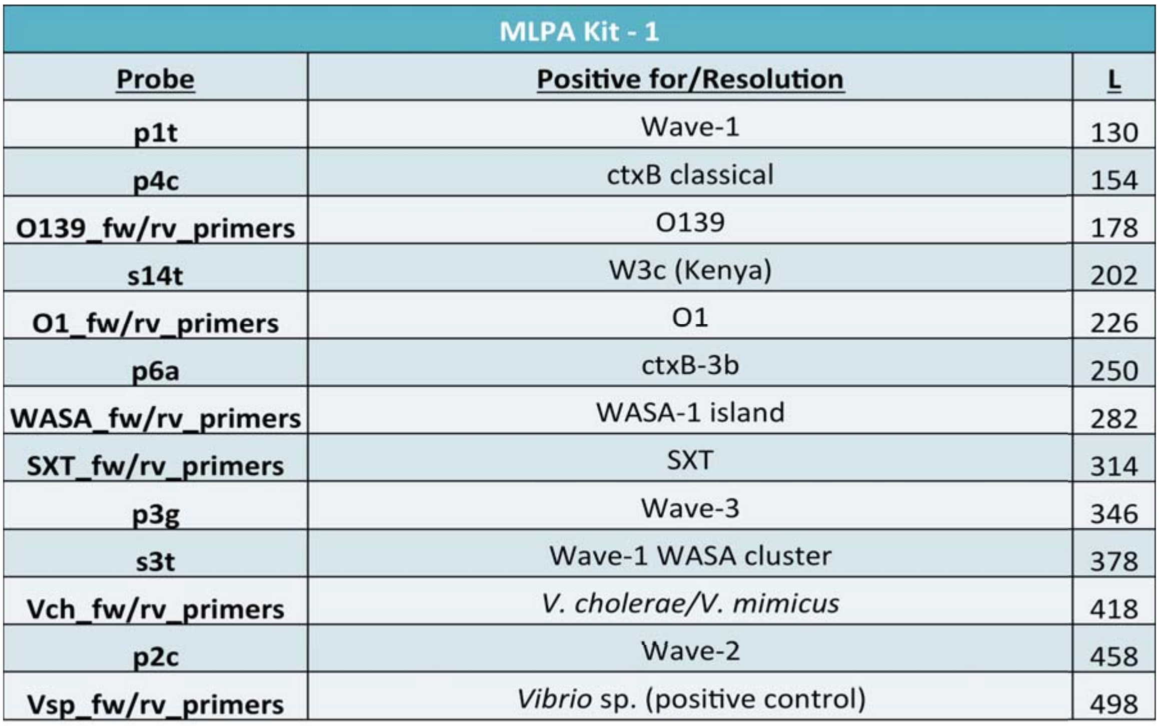

Haiti
North India
Nepal
East India and Bangladesh
Pakistan
3b
Nepal
East India and Bangladesh
Pakistan
Wave-3
3a
East India and Bangladesh
3c
Kenya
Zambia, Zimbabwe

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

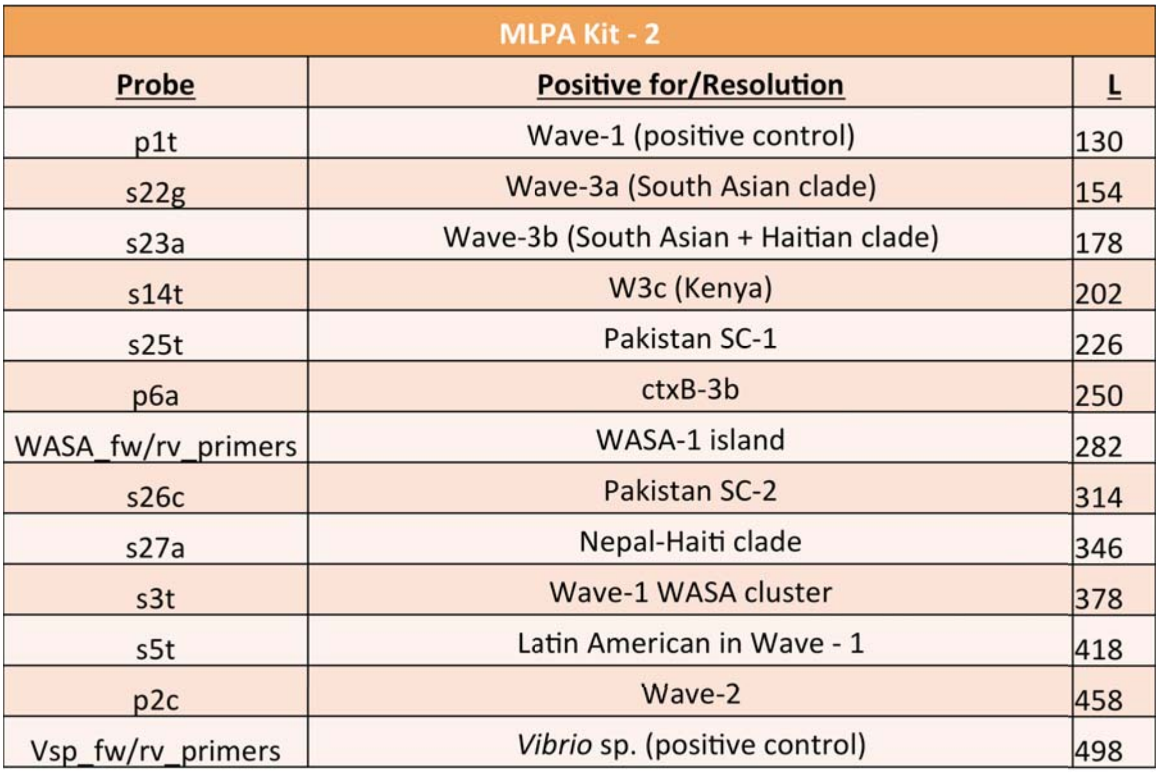
After the analysis based on the additional *V. cholerae* isolates it became clear that wave-3 could be sub-divided into three major sub-clades now referred to as 3a, 3b and 3c (Figure 5.5). The isolates in sub-clades 3a and 3b were primarily from the south-Asian sub-continent whereas isolates from Haiti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

5.2.2.2 Design of the MLPA based SNP-genotyping assays

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



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



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

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

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

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

The addition of new *V. cholerae* isolates to the seventh pandemic phylogeny consolidated the view that this is a highly monophyletic lineage. The linear regression analysis again showcased the clock-like evolution of the seventh pandemic *V. cholerae* O1 El Tor. When the new phylogenetic data was plotted onto the world map (Figure 5.6), the previously postulated pattern of spread of the seventh pandemic was conserved. All new clades and branches of the expanded phylgeny also radiated from the same source population, strengthening the previously proposed hypothesis that there is a single backbone population that is seeding cholera globally in the form of independent but overlapping waves that enter an area, cause cholera outbreaks that then die out when a new wave arrives. However, wave-1 *V. cholerae* that have apparently persisted in Mexico (chapter 3) are an exception to this hypothesis, and evidence for this persistence of this clade should be investigated across the regions of Latin American affected by the 1990s cholera outbreaks.

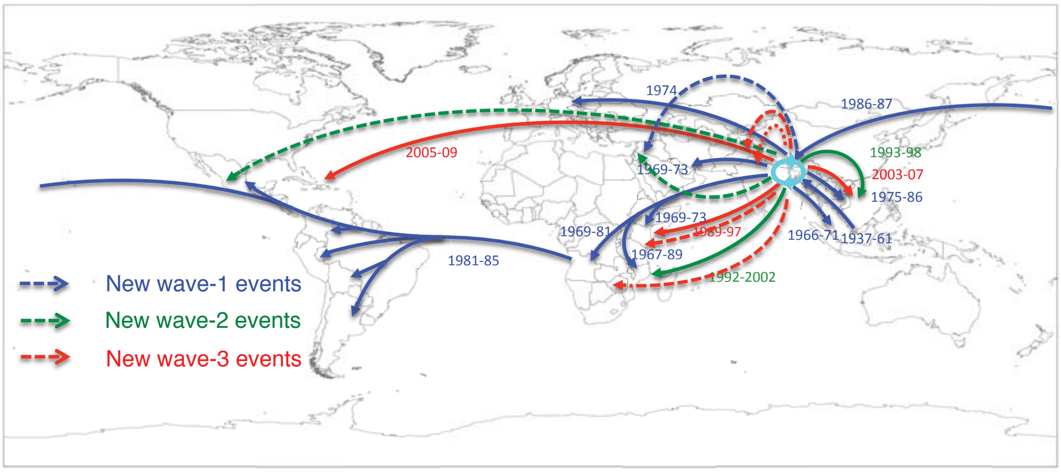


Figure 5.6: Figure showing the proposed spread of newly added *V. cholerae* isolates in the expanded seventh phylogenetic study. The dashed lines represent the newly plotted events and solid lines represent previously published data (Mutreja, *et al.*, 2011).

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

The traditional typing techniques used to define *V. cholerae* are not suitable for investigating the spread and epidemiology of cholera outbreaks and analysis based on whole genome sequences requires considerable resources and informatics skill. This is where the MLPA Kit based SNP genotyping approach described here could fill the gap and could provide robust answers based on the growing *V. cholerae* sequence databases. By using these kits, quick and reliable information about any isolates' position in the global seventh pandemic phylogenetic framework could be gained. This information could provide vital clues to facilitate tackling the spread of cholera. Even if an isolate cannot be mapped to the phylogeny using the proposed SNP genotyping, such a result could signify the emergence or identification of a new phylotype. In such an event, other public health centers would be alerted in advance, so their monitoring programs could look for such new clades and efforts to contain their spread could be initiated.