Chapter 3

Contrasting sources and behaviour of epidemic and endemic *Vibrio cholerae* in the Argentinian cholera epidemic, 1992-1998

Contribution statement

Nick Thomson, with Josefina Campos, supervised the work described in this chapter. My thesis committee - Julian Parkhill, Gordon Dougan, and George Salmond - contributed to the interpretation of these results. The isolate receipt data provided by INEI were collated and digitised principally by Tomás Poklepovich. The strain collection was curated by María Rosa Viñas and the *Enterobacterias* team at INEI-ANLIS "Dr. Carlos G. Malbrán", Buenos Aires, Argentina. DNA was isolated from *V. cholerae* by the *Enterobacterias* team, Daryl Domman and I in Argentina. I extracted DNA from samples at WSI with help from Charlotte Tolley.

I performed the analyses described in this chapter, and prepared all of the figures.

Publication

The results, figures, and text presented in this chapter have been submitted for publication.

Dorman MJ, Domman D, Poklepovich T, Tolley C, Zolezzi G, Kane L, Viñas MR, Panagópulo M, Moroni M, Binsztein N, Caffer MI, Clare S, Dougan G, Salmond GPC, Parkhill J, Campos J, Thomson NR. Genomics of the Argentinian cholera epidemic elucidate the contrasting dynamics of epidemic and endemic *Vibrio cholerae*. Revised and re-submitted (*Nature Communications*).

3.1 – Overview

An overall aim of this thesis is to compare the dynamics of pandemic *V. cholerae* to those non-pandemic *V. cholerae* which incur a lesser burden of disease, with a particular interest in describing variation in these distinct *V. cholerae* across time and geography, and at both molecular and phenotypic levels. As previously discussed (Introduction), although there is a clear difference between *V. cholerae* lineages in their ability to cause epidemics and pandemics, studies of gene content variation and even of SNVs have not yet elucidated the specific genetic determinants responsible for "pandemicity". However, these studies have been hampered by relying on relatively small numbers of laboratory strains, particularly of non-pandemic *V. cholerae*, and by a limited understanding of the population structure of the species as a whole.

One possible way to compare pandemic and non-pandemic *V. cholerae* is to examine differences in their evolutionary dynamics during an outbreak of pandemic cholera. Due to limited numbers of samples included in previous sequencing studies, we understand very little about how epidemic *V. cholerae* evolves once it is introduced into a population or a region, or how it contrasts to contemporaneous non-pandemic *V. cholerae* taken from the same region. The only study to date which has sequenced large numbers of closely geographically-related samples over a relatively short period of time has been performed in Bangladesh [235]. However, Bangladesh is hyper-endemic for cholera, and in that setting, multiple sub-lineages of 7PET were found to co-circulate with one another simultaneously. It is also known that there is background immunity to *V. cholerae* in this endemic setting [361]. It would be potentially misleading to infer from studies in such a setting rules that govern 7PET evolution and behaviour upon its introduction into a naïve population. Therefore, to study the evolutionary dynamics of a single 7PET sub-lineage, it was decided to re-visit the Latin American cholera epidemic of the 1990s.

In 1991, cholera returned to Latin America after an absence of nearly 100 years [173, 362]. Crucially, those epidemics which had occurred prior to 1991 had been associated with pandemics 2 through 5, all of which are believed to be due to Classical *V. cholerae* (Introduction, section 1.3.1.1; [36, 97, 162]). Therefore, the introduction of 7PET into Latin America in 1991 is an excellent example of point-source introduction of a pandemic bacterial clone into what can be considered an immunologically naïve population. Latin America

presented a unique opportunity to understand the longitudinal evolution of pandemic *V. cholerae* upon its introduction into a naïve population.

Epidemic cholera broke out in Peru in January 1991, and subsequently spread to Argentina, with the first cholera cases reported on the 5th of February 1992 in Salta province, close to the Bolivian border [174, 363, 364]. On 6th February 1992, a state of emergency was declared by Carlos Menem, then the Argentinian President, in response to outbreaks of cholera in towns in Salta province [365]. Thereafter, cholera cases were reported annually between 1992 and 1998 in Argentina [186, 366–368], with a cumulative total of 4,281 cases reported to PAHO and WHO for this period [363, 369–372].

The rapidity by which pandemic cholera can be transmitted is illustrated by an outbreak linked to Aerolineas Argentinas flight 386. On 14th February 1992, this flight departed Buenos Aires, Argentina, bound for Los Angeles. Following a stop-over in Lima, Peru [373], the flight landed in Los Angeles, carrying 336 passengers and 20 crew members [373, 374]. By 19th February 1992, *V. cholerae* serogroup O1 had been isolated from five passengers suffering from diarrhoea, and by 21st February, one elderly passenger had died from cholera [374–376]. The Argentinian government insisted that contaminated food taken on board during the stop-over in Lima had been responsible for the outbreak; conversely, the Peruvian government claimed that a passenger beginning their journey in Buenos Aires had been the source of the outbreak [377]. Simultaneously, reports of cholera were received by the California Department of Health Services and the Los Angeles County Department of Health Services [373, 374]. Of the 336 passengers on the flight, 100 were found to be carriers of *V. cholerae*, 75 reported diarrhoea of whom ten were hospitalised, one case of which was fatal [373]. Forty-eight *V. cholerae* isolates were obtained from these passengers, of which 34 were serotype Ogawa and 14 were serotype Inaba [373]. It is from this outbreak that the A1552 laboratory isolate, used for numerous functional studies of *V. cholerae* and the source of an important reference genome sequence used in this chapter, was obtained [378, 379].

Argentina is an ideal and unique setting in which to study the evolution of 7PET during the 1990s, because unlike some other countries in the region, the socioeconomic position of Argentina is thought to have enabled the monitoring and control of the epidemic [168, 380]. Argentina instituted mandatory notification of cholera cases nationwide during 1991 after cholera broke out in Peru [368], and developed public information campaigns which resulted

in a concomitant increase in the rate of diarrhoeal disease reporting [381]. During the epidemic period (1991-1998), all suspected cholera cases were tested in microbiology laboratories nationwide, and putative *V. cholerae* from clinical cases of suspected cholera, as defined by the Argentinian Ministry of Health, were isolated and sent to INEI-ANLIS "Dr. Carlos G. Malbrán" (INEI), the national reference laboratory of Argentina, for further confirmation. The INEI archives contain over 3,500 phenotypically-characterised *V. cholerae* isolates, which may represent over 82% of the WHO-reported cholera cases for the whole country from this 1990s epidemic (an epidemic which caused over 1.2 million disease cases of cholera across Latin America [168]).

In order to study the similarities and differences in evolutionary dynamics between epidemic and endemic *V. cholerae*, an appropriate strain collection is required. Previous genomic work looking across Latin America and including five Argentinian isolates [189, 234] suggested that a single toxigenic *V. cholerae* clone belonging to the LAT-1 sub-lineage of 7PET was responsible for the Argentinian cholera epidemic, related to that which caused outbreaks in Peru [189]. Therefore, we hypothesised that the comprehensive strain collection at INEI, together with the metadata recorded for each isolate, would enable the study of the progression of an epidemic attributed to one discrete introduction of 7PET.

Prior to the implementation of genomic analysis, observations made about cholera in Latin America have been difficult to resolve in the light of the cholera paradigm [29]. As discussed previously (Introduction, section 1.3.1.2), this paradigm suggested that epidemic cholera in Latin America was driven by environmental factors, which created an environment conducive to the expansion of local *V. cholerae* populations. These local *V. cholerae* were proposed to have been the aetiological agent of the 1990s epidemic [29, 382]. Although the role of the environment as a source of epidemic *V. cholerae* continues to be debated today [383, 384], the hypothesis that environmental sources gave rise to the Latin American cholera epidemics in 1991 has been shown to be incompatible with genomic data [189]. The Haitian cholera epidemic which began in 2010 has also been shown not to be derived from an environmental source [26, 189]. In addition, genomic evidence has shown that even toxigenic *V. cholerae* O1 that are local to Argentina [367], and to other Latin American countries, are distinct from the aetiological agent of the seventh pandemic cholera [189].

Previous *V. cholerae* sequencing projects have focused on studying isolates from clinical sources taken during outbreaks, and have therefore been enriched for serogroup O1 and toxigenic *V. cholerae* isolates. Indeed, it was only because of a combination of socioeconomic conditions and the materials preserved by certain laboratories that non-pandemic *V. cholerae* O1 lineages were sequenced in our previous study [189]. However, alongside the mandated collection of clinically-isolated *V. cholerae* O1, INEI also received and stored non-O1 *V. cholerae* of both clinical and environmental origin. Crucially, these non-O1 *V. cholerae* were isolated from the same times and places as the *V. cholerae* O1 during the epidemic. After the 1990s epidemic ended in Argentina, environmental surveillance projects saw additional non-O1 *V. cholerae* added to the INEI collection. Thus, it was reasoned that this collection of non-O1 isolates may represent the *V. cholerae* diversity present in, and endemic to, Argentina during and after the period of the cholera epidemic.

This thesis chapter describes an analysis of several hundred genomes of spatiotemporally diverse Argentinian 7PET isolates. These genomic data were then linked to related metadata and to historical literature, to describe in detail the dynamics and trajectory of the Argentinian cholera epidemic. Finally, a number of non-7PET *V. cholerae*, isolated during the time of the epidemic, were contrasted with the pandemic 7PET isolates, laying the foundation for more detailed analyses of non-7PET bacteria in subsequent chapters.

3.2 – Specific aims

This chapter aimed to use the INEI collection of *V. cholerae*:

a) To test the hypothesis of whether isolates belonging to the LAT-1 sub-lineage, which was introduced into South America in 1991, were the aetiological agents of epidemic cholera in Argentina between 1992 and 1998.
b) Using sequencing data, to study how epidemic *V. cholerae* evolved genomically over the course of the epidemic (seven years).
c) To characterise the genomes of non-O1 (and, presumably, non-epidemic) *V. cholerae*, and contrast the genomes of these isolates with those of contemporaneously-isolated pandemic *V. cholerae*.

3.3 – Ethical statement

The isolates described herein were acquired by the national reference laboratories at INEI-ANLIS as part of the routine receipt of notifiable pathogens during this epidemic. No patient data, identifiable or otherwise, were made available for the analyses reported in this thesis. Accordingly, ethical approval was not required.

3.4 – Results

3.4.1 – WHO/PAHO records

In order to obtain an initial insight into the dynamics of the Argentinian cholera epidemic, the number of WHO and PAHO-reported cholera cases for the country were plotted (Figure 3.1). These data are cumulative, and report the number of cases of disease *per annum*.

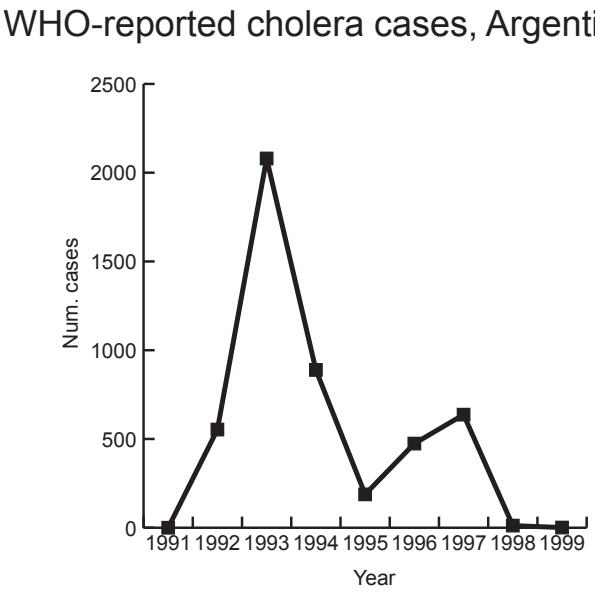

Figure 3.1 – Cholera cases reported to PAHO and WHO from Argentina, 1991-1999. Data were taken from [363, 369–372, 385, 386].

These data suggest that there were two 'peaks' of cholera incidence in Argentina between 1992 and 1998, one in 1993 (2,080 cases) and one in 1996/1997 (474 and 637 cases respectively) (Figure 3.1). However, there are discrepancies between these apparent maxima and those presented in other governmental reports and research publications, which have suggested that there were 'seven epidemics' of cholera in Argentina between 1992 and 1997 [186, 366–368].

3.4.2 – The records of isolate receipt at INEI-ANLIS

All *V. cholerae* received by INEI since 1992 have been recorded and documented, and these records were made available for this PhD thesis. Dates of isolate submission to INEI were available for each isolate, as was the location from which each *V. cholerae* was obtained, the serogroup, and for *V. cholerae* O1, the serotype. GPS co-ordinates were determined for each

province or city from which *V. cholerae* had been received at INEI, and these were used to produce maps of *V. cholerae* O1 and non-O1 receipt over time (Figure 3.2). These were therefore used as a proxy for identifying the location and magnitude of disease incidence across the country (Figure 3.2).

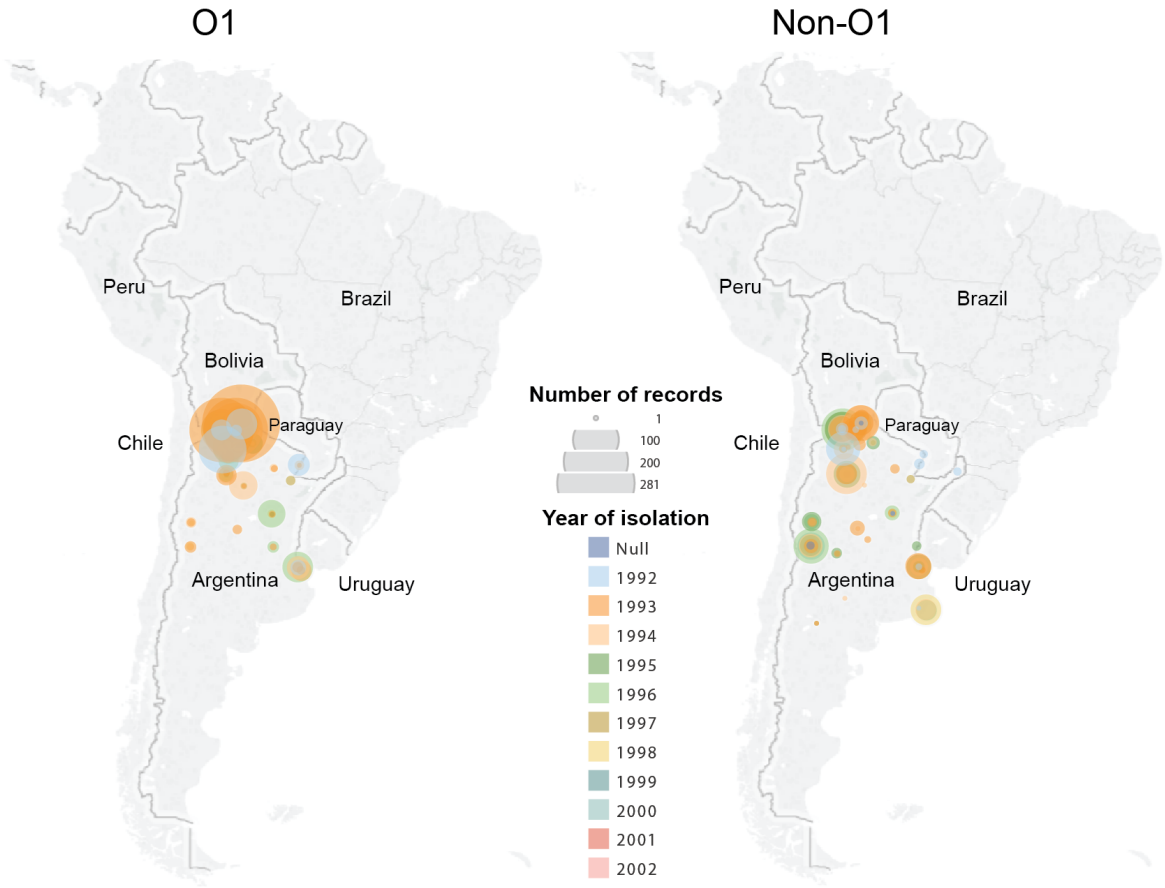

Figure 3.2 – Origins of *V. cholerae* received by INEI, 1992-2002. The non-O1 isolates obtained from coastal regions south of Buenos Aires were obtained from environmental surveillance projects at Instituto Nacional de Investigación y Desarrollo Pesquero (INIDEP), Mar del Plata. The size of the circle scales with the number of isolates received from a particular region in the year indicated.

V. cholerae were received from the North and Centre of Argentina. The geographic area covered by these isolates is approximately 1.2 million km^2 (calculating the area of Argentina North of a latitude line from Buenos Aires). The nature of the national reporting scheme in Argentina at the time, enforced at the national level, made it mandatory that suspected cases of

cholera were tested for *V. cholerae* nationwide. This meant that the absence of isolates from the South represents a true absence of cholera reports and isolation from this region.

The date associated with each bacterial isolate corresponded to the date on which the isolate was acquired by the reporting laboratories nationwide, rather than the date on which the isolate was received by INEI in Buenos Aires. The precise date of isolation was recorded for each isolate. Accordingly, it was reasoned that these data might describe the temporal variation in *V. cholerae* reported in Argentina in greater detail than the PAHO/WHO statistics (Figure 3.1). Dates of isolation were plotted across the period of the epidemic and beyond, to account for environmental sampling that continued beyond the end of the epidemic in 1998 (Figure 3.3).

Of the isolates received, 2,189 were serogroup O1 (60.2 %), and 1,308 were non-O1 *V. cholerae*. Clinical isolates made up the majority (2,077, 94.8%) of *V. cholerae* O1; 112 isolates were either environmental isolates or their sources were not recorded. Of the non-O1 isolates, 714 were of clinical origin (54.5%). There were 134 isolates for which there were no serogroup data recorded (n = 129), that were autoagglutinable (n = 4), or were recorded as being of serogroup O139 (n = 1). Of these 134 isolates, 106 were isolated in January 1993 and correspond to a reported outbreak of *V. cholerae* O1 (Josefina Campos, personal communication). Serotype data were not recorded for 25 of the *V. cholerae* O1 in these records (1.1%).

At least six peaks of *V. cholerae* O1 receipt can be observed in Figure 3.3. A seventh peak may be visible in early 1998. These are consistent with reports of 'seven epidemics' of cholera in Argentina during the 1990s [186, 366–368]. Although there were periods during which no *V. cholerae* O1 were received by INEI (Figure 3.3), non-O1 *V. cholerae* were received more consistently during the 1990s, though their receipt rose coincidentally with peaks in *V. cholerae* O1 receipt. The apparent discordance between these data and the PAHO/WHO records likely reflects the fact that the PAHO/WHO data are only available as annual case/fatality statistics, and are not broken down by month.

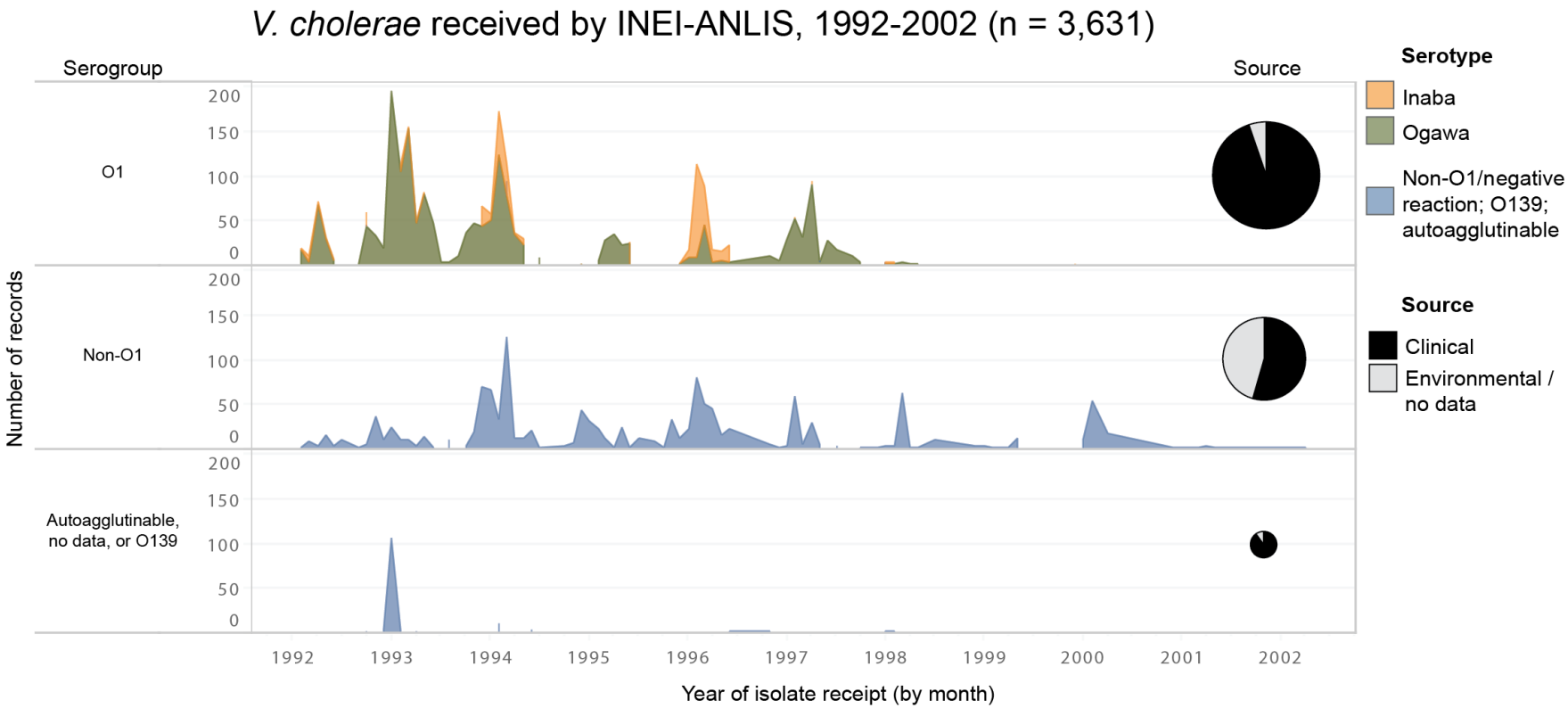

Figure 3.3 – Dates of isolation for *V. cholerae* received by INEI, 1992-2002. X-axis intervals are months of isolation. *V. cholerae* O1 are coloured by serotype. The fraction of clinical to environmental isolates are presented in pie charts (exact numbers are reported in section 3.4.2).

Consistent with a previous report that *V. cholerae* Ogawa was predominant in Argentina during the 1990s epidemic [387], Ogawa isolates accounted for 82.0% of the *V. cholerae* O1 in these records (1,795 isolates), and 369 isolates were serotype Inaba (16.8%). This is despite the initial 1991 Latin American cholera epidemic being ascribed to *V. cholerae* Inaba [369].

3.4.3 – Selection of isolates to sequence for this study

The isolates that were used for this project had been selected in order to represent as much of the country as possible (Figure 3.4) and to span the beginning and the end of the epidemic (Figure 3.5). A strong bias was applied towards sequencing *V. cholerae* O1, from 1992 and 1996/7 and across Argentina, in order to allow questions to be asked about how epidemic *V. cholerae* evolved across space and time during the epidemic. Contemporaneous non-O1 *V. cholerae* from the same geographical regions were also sequenced, in order to compare and contrast the genomes of these non-O1 (and presumably, non-epidemic) bacteria with those of O1 isolates.

Figure 3.4 – Locations from which the isolates analysed in this study were obtained. Of the 490 analysed isolates, 475 had region of origin recorded (96.9%). These data are not stratified by serogroup.

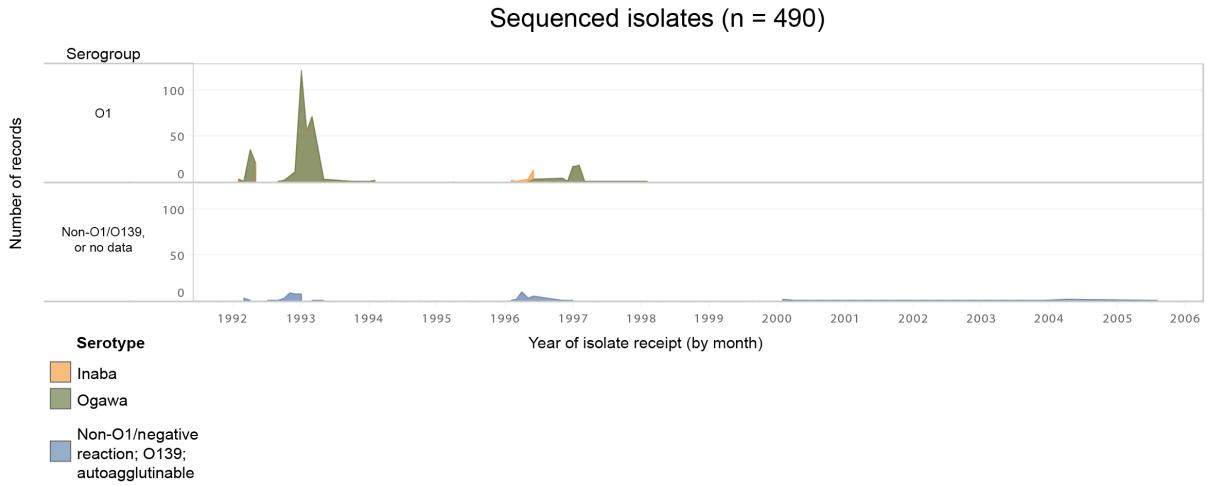

Figure 3.5 – Dates of isolation for the bacteria sequenced and analysed in this study. Dates of isolation were not recorded for four of the sequenced isolates (486/490 isolates represented in this figure; 99.1%).

Genomic DNA was extracted from a total of 511 *V. cholerae* stored at INEI and sequenced on the Illumina X10 platform for the purpose of this PhD project. Twenty-one of the resultant sequences were excluded after failing to meet quality control criteria (see Methods). The distribution of assembly lengths before the application of a 5 Mbp threshold is presented in Figure 3.6 (the *V. cholerae* genome is approximately 4.1 Mbp in length [59]). Kraken reports were also used to identify and remove sequences which were likely to be heavily contaminated with non-*V. cholerae* DNA, or sequences corresponding to species other than *V. cholerae*, as were preliminary phylogenetic trees which were used to exclude isolates on extremely long branches (data not shown).

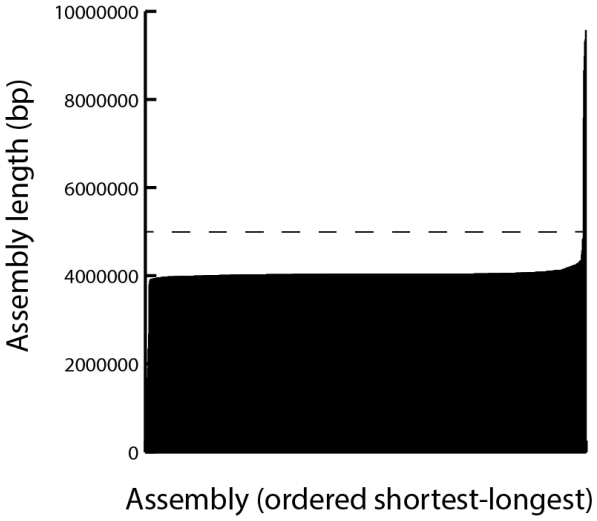

Figure 3.6 – Example of the application of an assembly length cut-off to SPAdes assemblies produced in this study. The dashed line corresponds to an assembly length of 5 Mbp – sequences of a longer assembly were assumed to be contaminated or of poor quality, and were excluded from further analysis. These 936 assemblies include previously-published sequences.

Once low-quality sequences and contaminated samples were excluded, this left a total of 490 genome sequences which were used for all subsequent analysis. A core-gene phylogeny was calculated using the 490 Argentinian sequences, together with a set of 7PET and non-7PET genomes (1,165 total genomes), in order to classify the Argentinian isolates as 7PET, or non-7PET, based on their phylogenetic position (Figure 3.7). Based on these data, 425 of the 490 sequenced *V. cholerae* were determined to be members of 7PET (86.7%) and 65 were classified as non-7PET (Figure 3.7).

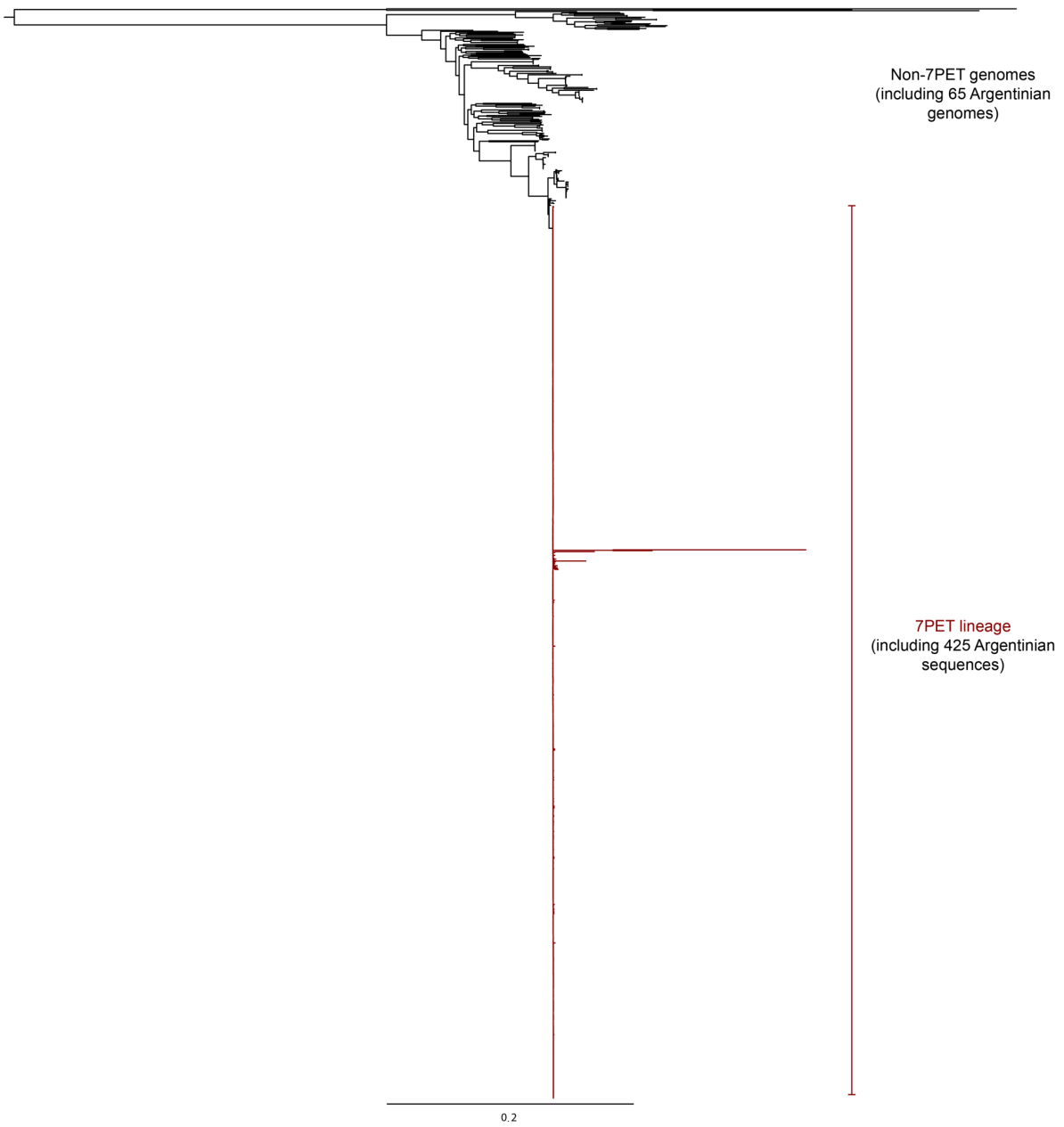

Figure 3.7 – Maximum-likelihood core-gene phylogeny of 1,165 *V. cholerae*. Isolates sequenced in this study (n = 490) were determined to be 7PET, or not, on the basis of their phylogenetic position. The phylogeny was calculated using FastTree v2.1.10 [388] and used as an intermediate analysis step.

3.4.4 – 7PET phylogeny

The 425 7PET genomes were placed into phylogenetic context alongside 517 additional genomes [189], by mapping the reads for these sequences to the *V. cholerae* reference sequence (strain N16961 [59]) and calling non-recombinant SNVs using established analysis pipelines (see Methods). The sequences of both *V. cholerae* chromosomes in the N16961 reference were combined for this purpose. Within the resultant alignment of 942 sequences, Gubbins masked

11.83% of the genome as being potentially recombined. When regions of recombination that were associated solely with pre-pandemic strain M66 were removed from consideration, 2.4% of the genome was predicted to be recombined (Figure 3.8).

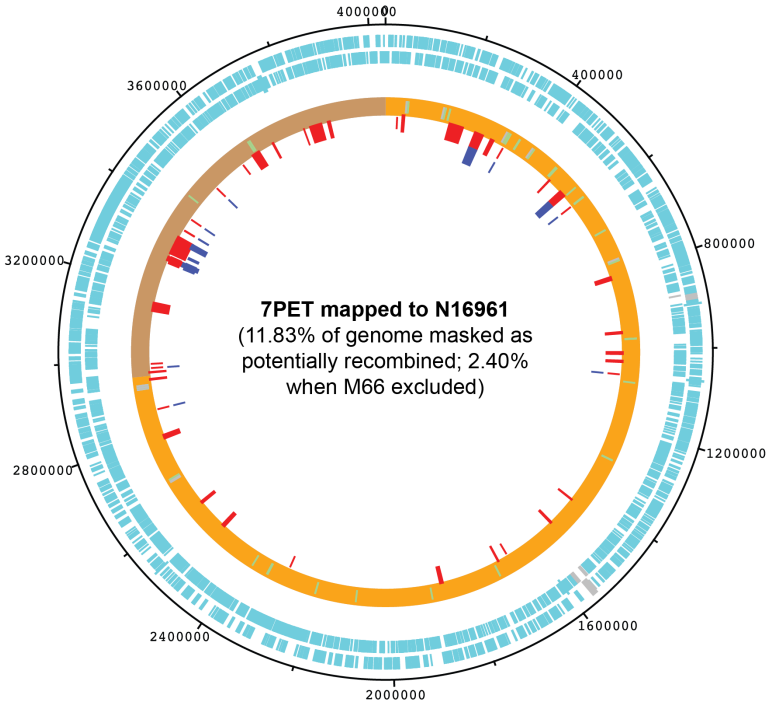

Figure 3.8 – N16961 chromosome regions predicted to be recombined from the 7PET alignment. Regions predicted by Gubbins to be recombined using the alignment of 942 *V. cholerae* sequences are indicated in red. The regions that are not solely associated with the pre-pandemic isolate M66 are indicated in blue (innermost ring). The two outermost rings indicate the presence of open reading frames on the forward and reverse strand, respectively. The sequences of both N16961 chromosomes were concatenated to produce this figure.

A maximum-likelihood phylogenetic tree was then calculated from the non-recombinant SNVs in this alignment (Figure 3.9). Parsimony-informative SNVs were identified from the alignment and used to cluster the data with Fastbaps [334]. These clusters were used to guide the assignment of 7PET isolates to sub-lineages (Figure 3.9). Of the 425 7PET isolates from Argentina, 421 isolates were members of the LAT-1 sub-lineage, which had been introduced to Peru in 1991 and subsequently spread across Latin America [189]. Since none of the Argentinian genomes were members of any other 7PET sub-lineage, the hypothesis that epidemic cholera in Argentina was caused by the same strain introduced into Peru in 1991 is very strongly supported.

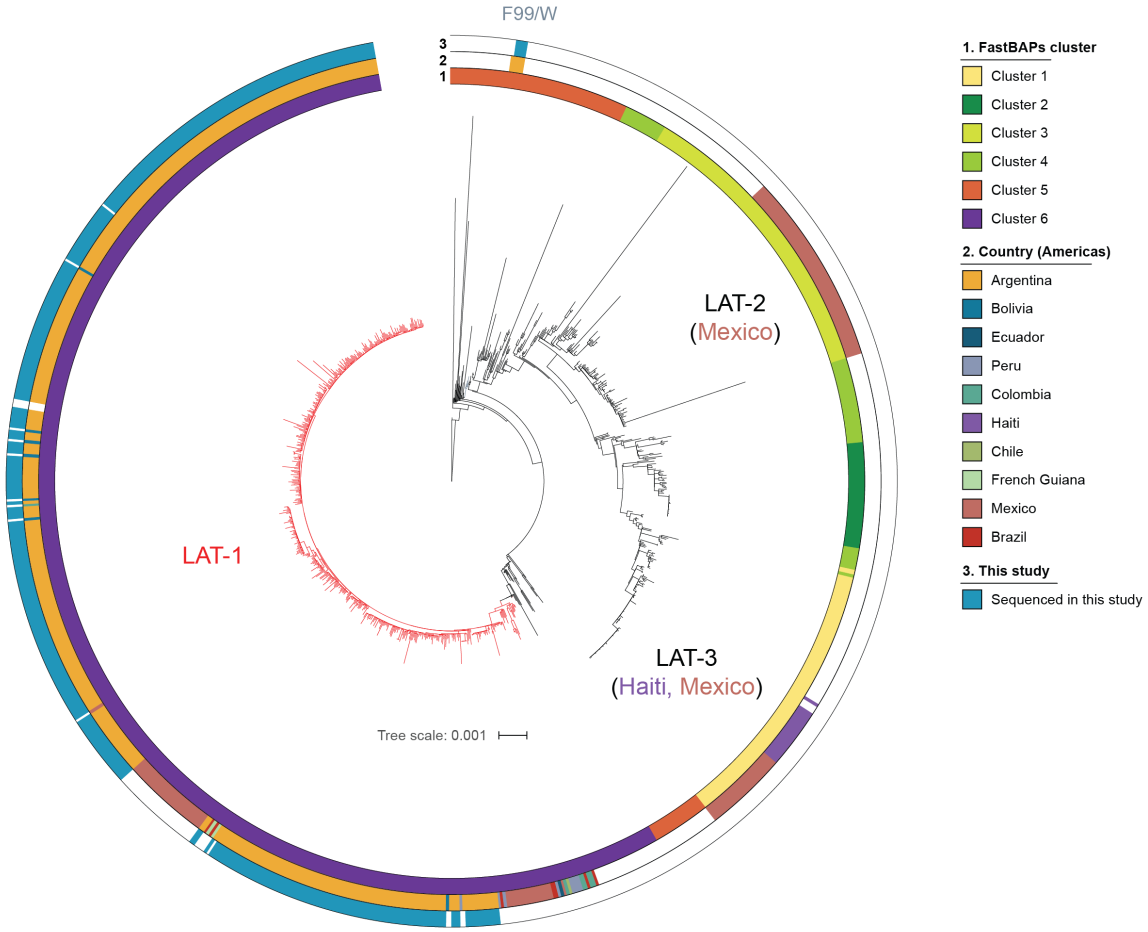

Figure 3.9 – A maximum-likelihood phylogeny of 7PET. Tree calculated using 7,556 non-recombinant SNVs, and rooted on the pre-pandemic isolate M66. Clustering was performed using Fastbaps and an alignment of 3,874 parsimony-informative SNVs. Countries of origin for sequences from South and Central America are reported. LAT transmission events as described previously [189] are indicated; LAT-1 is indicated in red.

The remaining four isolates were not members of LAT-1, but were phylogenetically closely related to F99/W (Figure 3.9). This is a isolate that had been reported previously [189, 389] and had been shown to be neither toxigenic nor a member of LAT-1 [189]. The genomes of the four isolates related to F99/W were investigated in more detail using comparative genomics, to determine the presence and absence of pathogenicity islands associated with 7PET [54, 133] (Table 3.1). All but one were found to encode VPI-1, the genomic island which encodes TCP, the receptor for CTXφ [55]. None of these isolates were found to carry CTXφ or the *ctxAB* genes required to express CT, consistent with the original report of the F99/W genome sequence [189].

Pathogenicity island	F99/W isolate				
	16853_4#67 (F99/W)	26268_1#261 (CCBT0194)	26268_1#259 (CCBT0192)	26268_1#257 (CCBT0190)	26268_1#254 (CCBT0187)
VPI-1	Yes	No	Yes	Yes	Yes
VPI-2	Yes	Yes	Yes	Yes	Yes
VSP-1	Yes	Yes	Yes	Yes	Yes
VSP-2	Yes	Yes	No	Yes	Yes
CTXφ	No	No	No	No	No

Table 3.1 – Presence of select pathogenicity islands in F99/W isolate genomes.

ACT comparisons were used to confirm that these pathogenicity islands are present, and absent, in their totality from the canonical integration sites (based on the N16961 reference genome; Figures 3.10 to 3.12). The absence of CTXφ was also verified using pangenome gene presence/absence data in subsequent analyses, to exclude the possibility of the bacteriophage integrating into non-canonical loci. Similarly, in the case of CCBT0194, mapping data (against N16961) were visualised in order to verify the absence of VPI-1, by confirming that these regions in the reference genome were not covered by mapped reads from the relevant sequencing run (Figure 3.11).

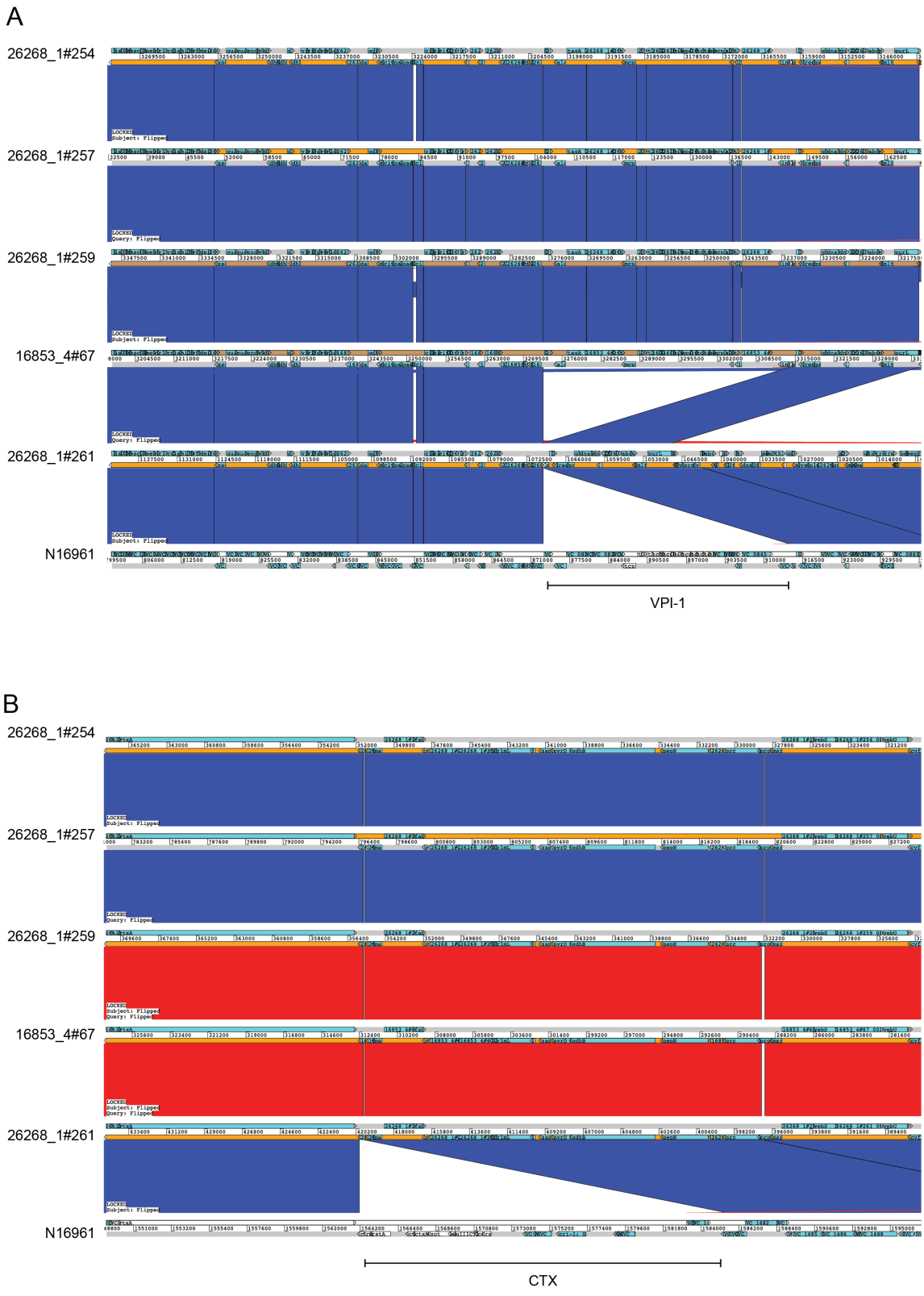

Figure 3.10 – Confirming the presence and absence of VPI-1 and CTXφ in isolates phylogenetically-related to F99/W. (A): The sequence corresponding to VPI-1 is absent from CCBT0194 (26268_1#261) but present in all other sequences being compared, including the N16961 reference. (B): The CTXφ prophage is absent from the canonical integration locus on the larger chromosome in each of the isolates in the F99/W clade (Figure 3.9).

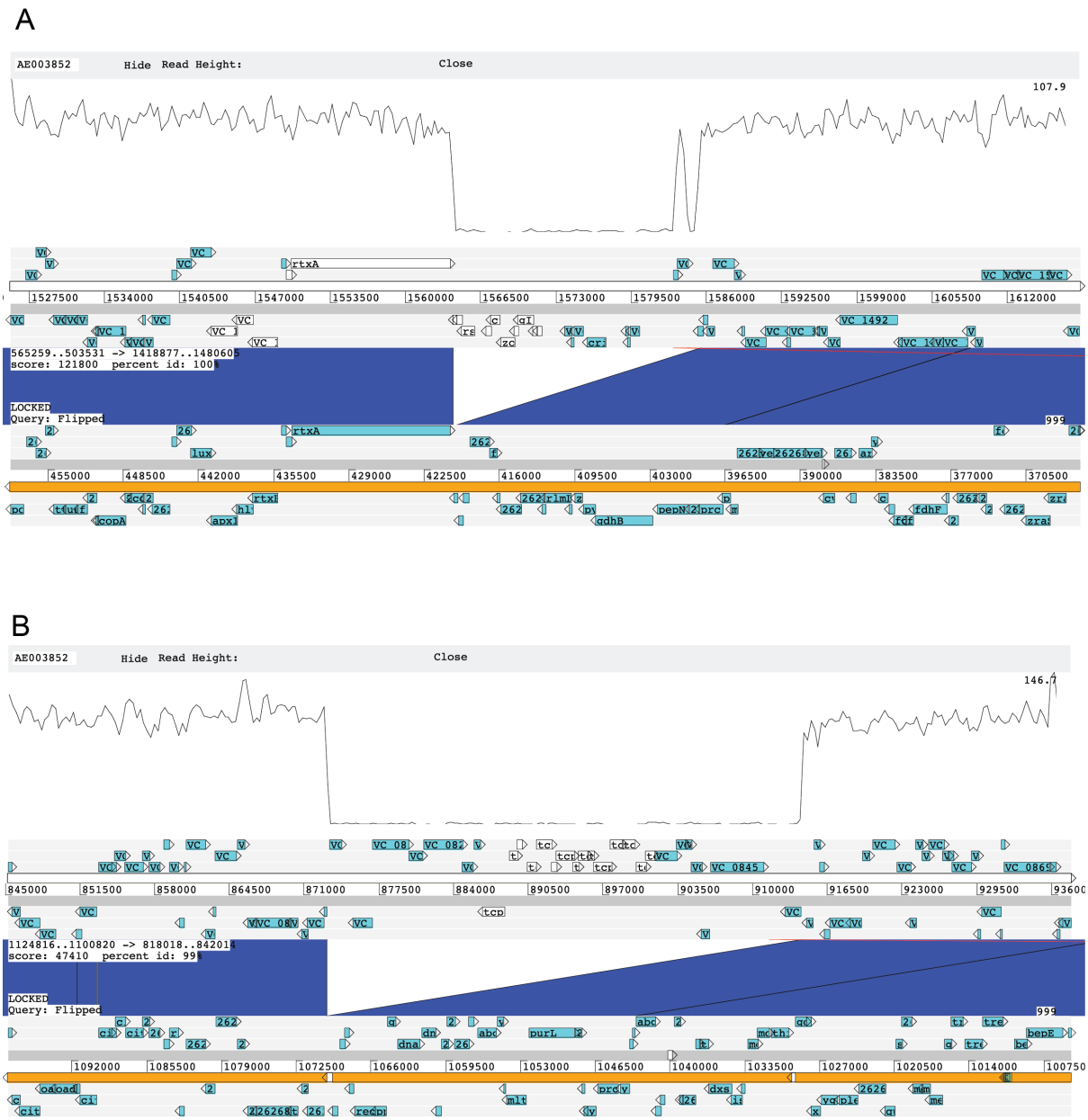

Figure 3.11 – Confirming the absence of CTXφ and VPI-1 from CCBT0194. The depth of reads mapped to the N16961 reference from isolate CCBT0194, the N16961 sequence, and the CCBT0194 assembly are presented. Panel (A) shows the section of the chromosome that harbours the CTXφ prophage and panel (B) shows the VPI-1 integration locus. The drop in coverage to near-zero over these pathogenicity islands, together with the absence of these islands from the CCBT0194 assembly, supports the conclusion that this isolate lacks both CTXφ and VPI-1.

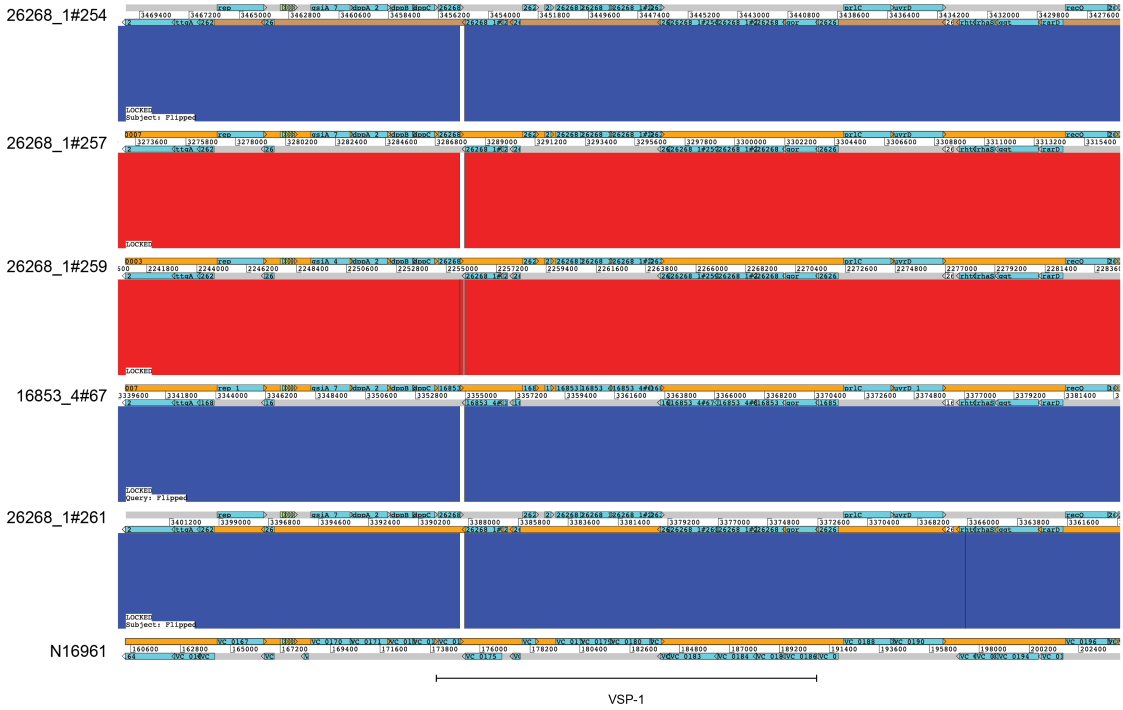

Figure 3.12 – Confirming the presence of VSP-1 in F99/W isolate genomes. VSP-1 is present in its canonical integration locus in all isolates in the F99/W clade (Figure 3.9).

Since the F99/W clade of isolates harbour chromosomal pathogenicity islands that are associated with pandemic *V. cholerae* (Table 3.1), including VPI-1, it is interesting to speculate as to whether these have the potential to be infected by the CTXφ bacteriophage and to become toxigenic. However, live cultures of these isolates were not available with which to test this hypothesis.

3.4.5 – LAT-1 phylogenetics

Although all of the Argentinian LAT-1 genomes clustered together, there appeared to be additional topological structure within the LAT-1 phylogeny (Figure 3.9). Therefore, to study this in greater detail, we re-mapped the reads for the 531 sequences in the LAT-1 sub-lineage to the closed genome sequence of strain A1552 [326], and called SNVs against this reference. A1552 was obtained by the Schoolnik laboratory from the Californian Department of Health Services [379, 390]. It is an Inaba El Tor bacterium [326, 378], and was isolated from a traveller

suffering from cholera in California in 1992 who arrived on Aerolineas Argentinas flight 386 (section 3.1; [378, 379]).

Based on this alignment of 532 sequences, just 0.03% of the A1552 genome was predicted to be recombined (Figure 3.13). A maximum-likelihood phylogeny was calculated using an alignment of 2,651 non-recombinant SNVs (Figure 3.14A). A mean of only 26.05 non-recombinant SNVs across both chromosomes separated the sequence of each LAT-1 isolate from that of the A1552 reference genome (min 10, max 149, stdev 14.10; Figure 3.14B).

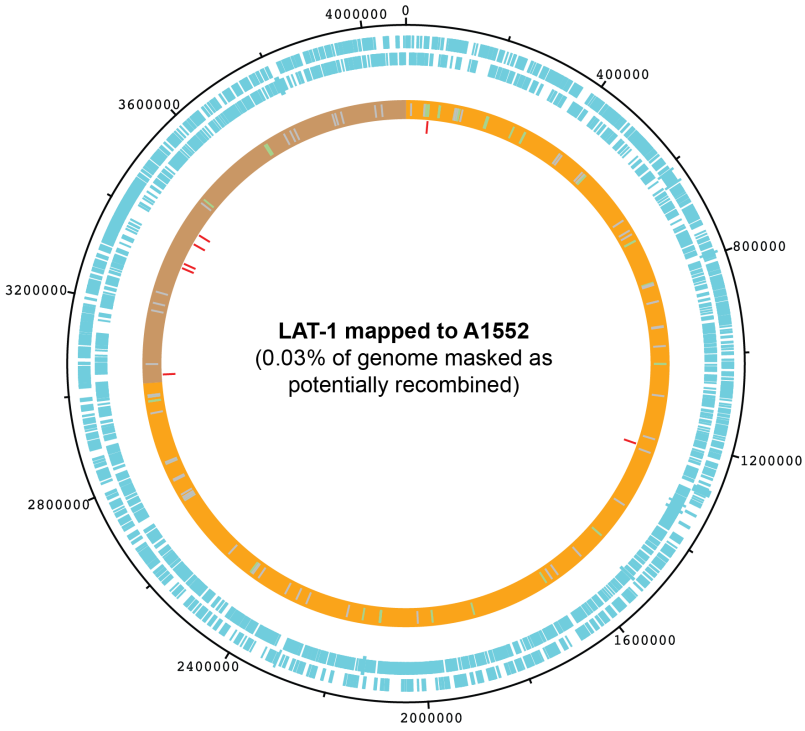

Figure 3.13 – A1552 genome regions predicted to be recombined using the LAT-1 dataset. The sequences of both the larger and smaller chromosomes were concatenated to produce this figure (orange, brown respectively). The inner ring (red) indicates the sections of the genome sequence predicted to be recombined by Gubbins. tRNAs are indicated by green ticks. The two outermost rings indicate the presence of open reading frames on the forward and reverse strand, respectively.

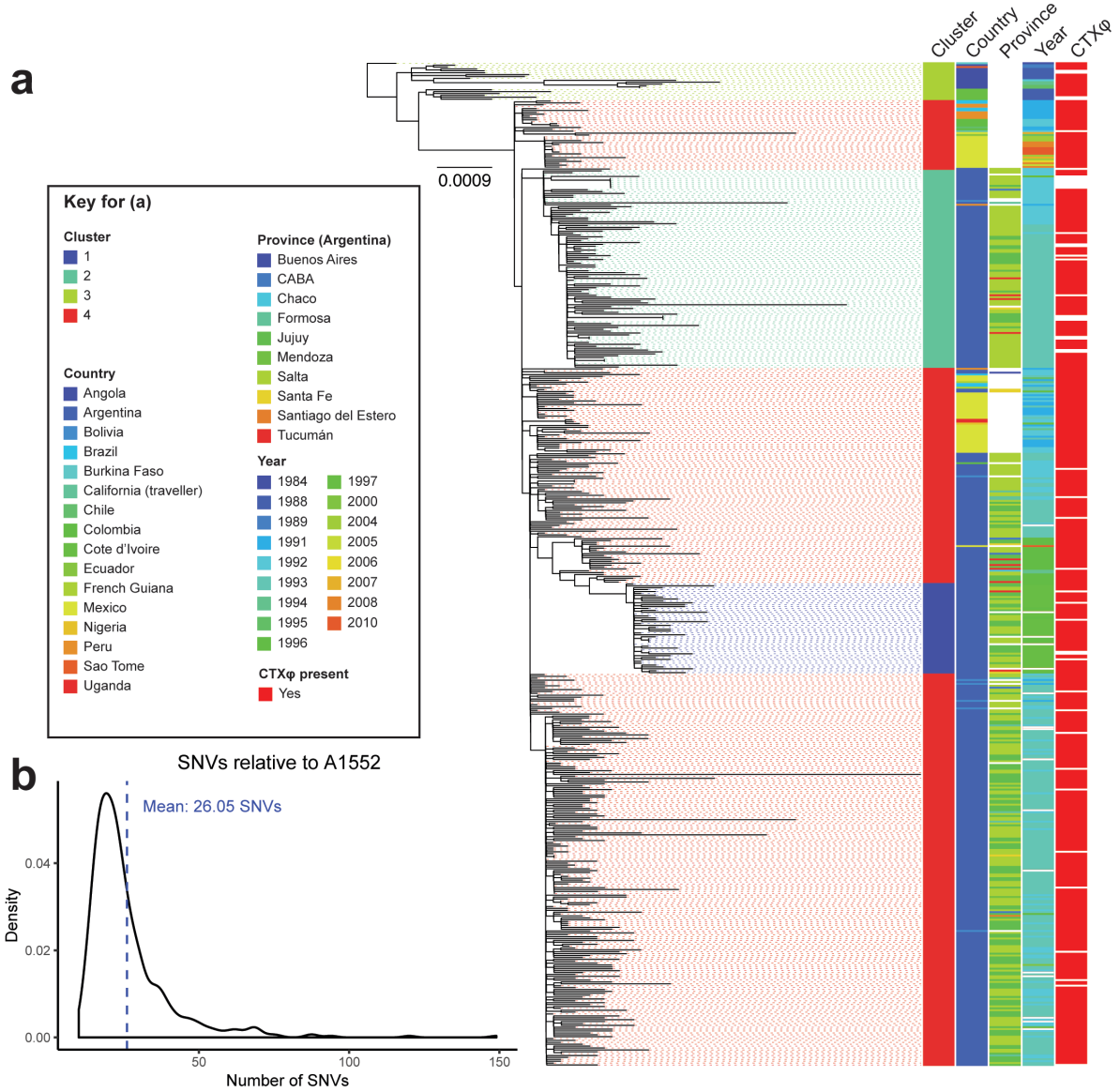

Figure 3.14 – LAT-1 phylogeny and SNV distances from A1552. (A): A maximum-likelihood phylogeny calculated from an alignment of 2,651 non-recombinant SNVs relative to the A1552 reference (accession # CP025936, CP025937). The tree is rooted using the genome of an isolate from Burkina Faso taken in 1984, CNRVC980048. This isolate lacks the WASA-1 genomic marker of LAT-1 [158, 189, 234]. Metadata and the presence of the CTXφ prophage are indicated. An alignment of 725 parsimony-informative non-private SNVs was used to cluster the data using Fastbaps. (B): A density plot of non-recombinant SNVs separating each genome from A1552.

From Figure 3.14 it is clear that LAT-1 sequences did not cluster by geography, either by province or region (Figure 3.14A). Isolates from different Northern provinces were interspersed amongst one another, as were isolates from countries bordering Northern Argentina such as Bolivia. However, limited clustering by date of isolation could be observed. Argentinian isolates from 1996 and 1997 clustered together, and cluster 1 was almost fully

made up of isolates from 1997. Cluster 2 contains Argentinian isolates from 1992 and 1993, and one isolate from 1997, from multiple provinces. Additionally, cluster 2 contained one Bolivian genome from 1992 and a Peruvian genome from 1991.

Consideration was given to whether a dated phylogeny could be used to study transmission of LAT-1 within Argentina. However, a robust temporal signal was not detected in these data – plotting root-to-tip divergence of the LAT-1 phylogeny (Figure 3.14A) against time using TempEst v1.5.1 yielded a regression with a poor correlation coefficient and R^2 value (Figure 3.15). It was also noted that of the 2,651 SNVs in the LAT-1 alignment, 1,926 (72.6%) were private to single genomes in the dataset. This strongly suggested that each isolate may have evolved privately during storage. This may be a consequence of the way in which the isolates were stored; some isolates were stored for up to 27 years before being sequenced, and many had been preserved on stabs or at ambient temperature – as has been described previously, these are storage conditions that can select for hypermutator phenotypes, which would be consistent with accelerated private mutation [308, 391]. Therefore, to avoid drawing conclusions based on potentially spurious results, it was decided that it was inappropriate to calculate dated phylogenies using these data.

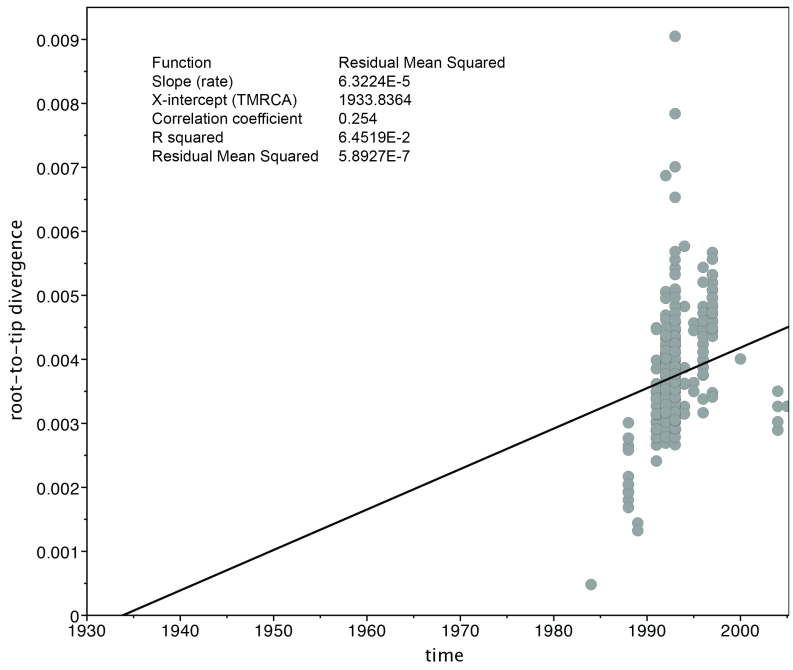

Figure 3.15 – Root-to-tip distance *versus* time for the LAT-1 phylogeny. Argentinian sequences for which dates of isolation were not recorded (n = 4) were arbitrarily assigned to the year 1993. Figure produced using TempEst v1.5.1.

A pangenome was calculated using these 532 LAT-1 genomes. A total of 3,412 core genes were identified from a total of 6,860 genes in the pangenome (core: 97% <= strains <= 100%, see Methods). The gene presence/absence matrix was interrogated to determine the presence and absence of WASA-1, *ctxAB*, and CTXφ across the dataset, and to detect genes associated with IncA/C plasmids (Figure 3.16).

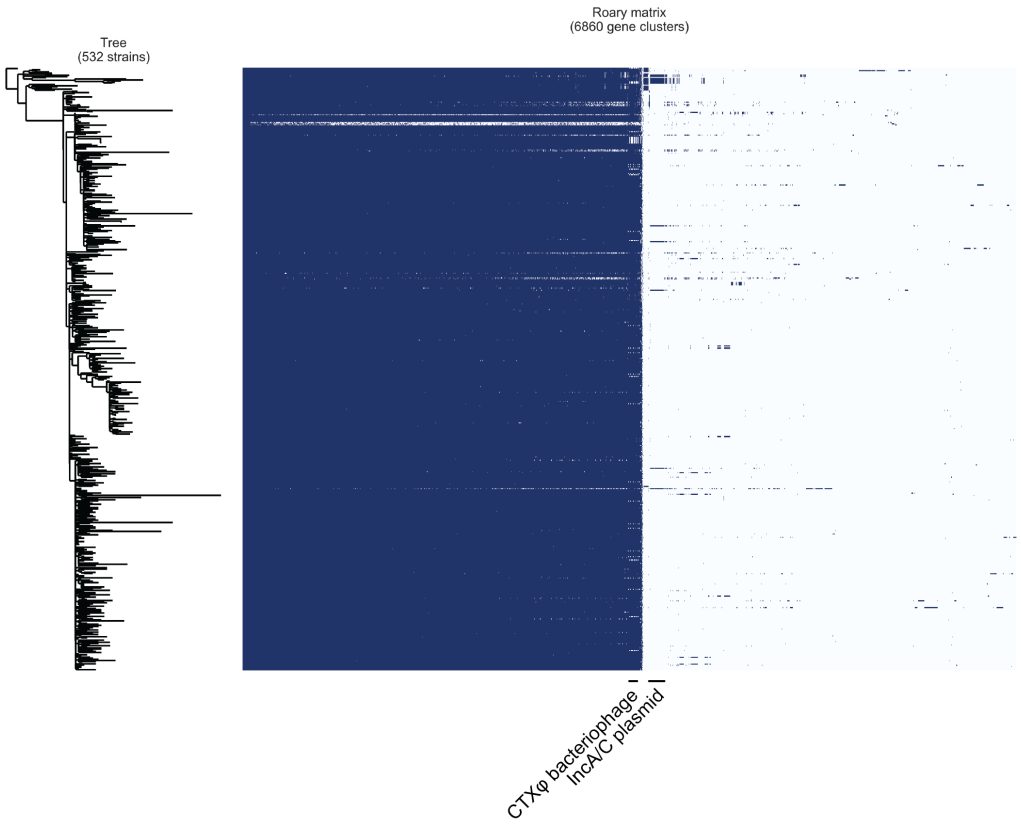

Figure 3.16 – LAT-1 pangenome gene presence/absence matrix visualisation. Phylogeny is as presented in Figure 3.14A. The position of CTXφ bacteriophage genes and those genes present on IncA/C plasmids are highlighted. White horizontal lines in the core genome block correspond to a small number of poorly-assembled sequences, which have been included to contextualise these sequences with previously-published studies.

In addition to the small accessory genome in these LAT-1 isolates, it was observed that the addition of new genomes to this analysis contributed negligible numbers of new genes to the pangenome overall – after the addition of 532 sequences, just ~1,800 unique genes had been added (Figure 3.17). This contrasts strongly with that observed in the non-LAT-1 genomes (see section 3.4.10).

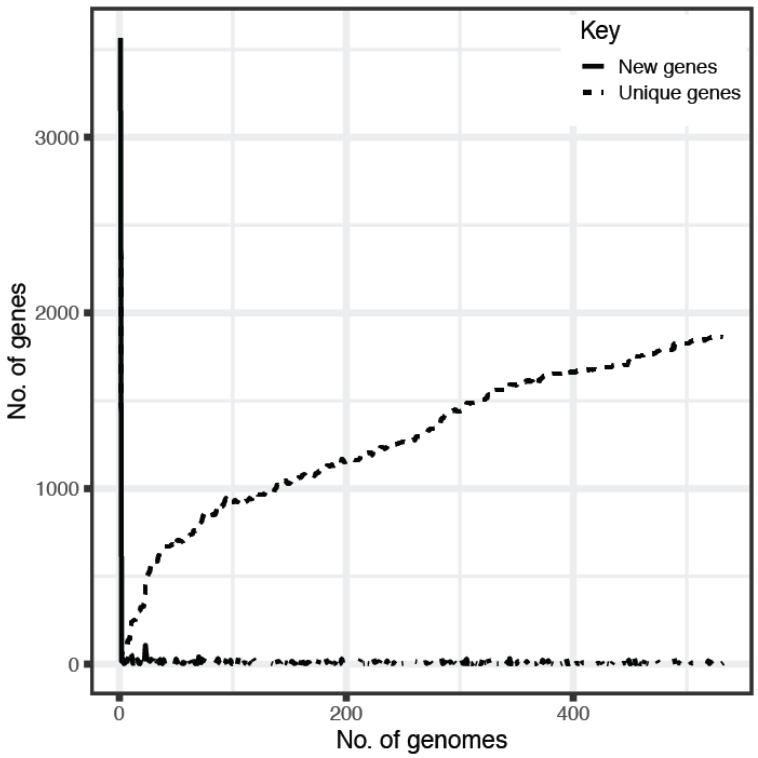

Figure 3.17 – Gene discovery as LAT-1 genomes are added to the pangenome. The initial peak of new gene detection corresponds to the identification of the core genome.

The presence of WASA-1 within LAT-1, a genomic island serving as a marker of the lineage [234], was confirmed by extracting the complete nucleotide sequence of WASA-1 from the A1552 genome sequence and using it to query each LAT-1 assembly with BLASTn. In order to confirm the presence of *ctxAB* and CTXφ, and to detect antimicrobial resistance genes and plasmid replicons, ARIBA was used together with the ResFinder and PlasmidFinder databases, and a custom database of *V. cholerae* virulence genes (Figure 3.18). ARIBA was also used to identify variation in the sequence of the *wbeT* gene, mutations in which are responsible for the switch from Ogawa to Inaba serotype [197, 198] (Introduction, section 1.3.1.3).

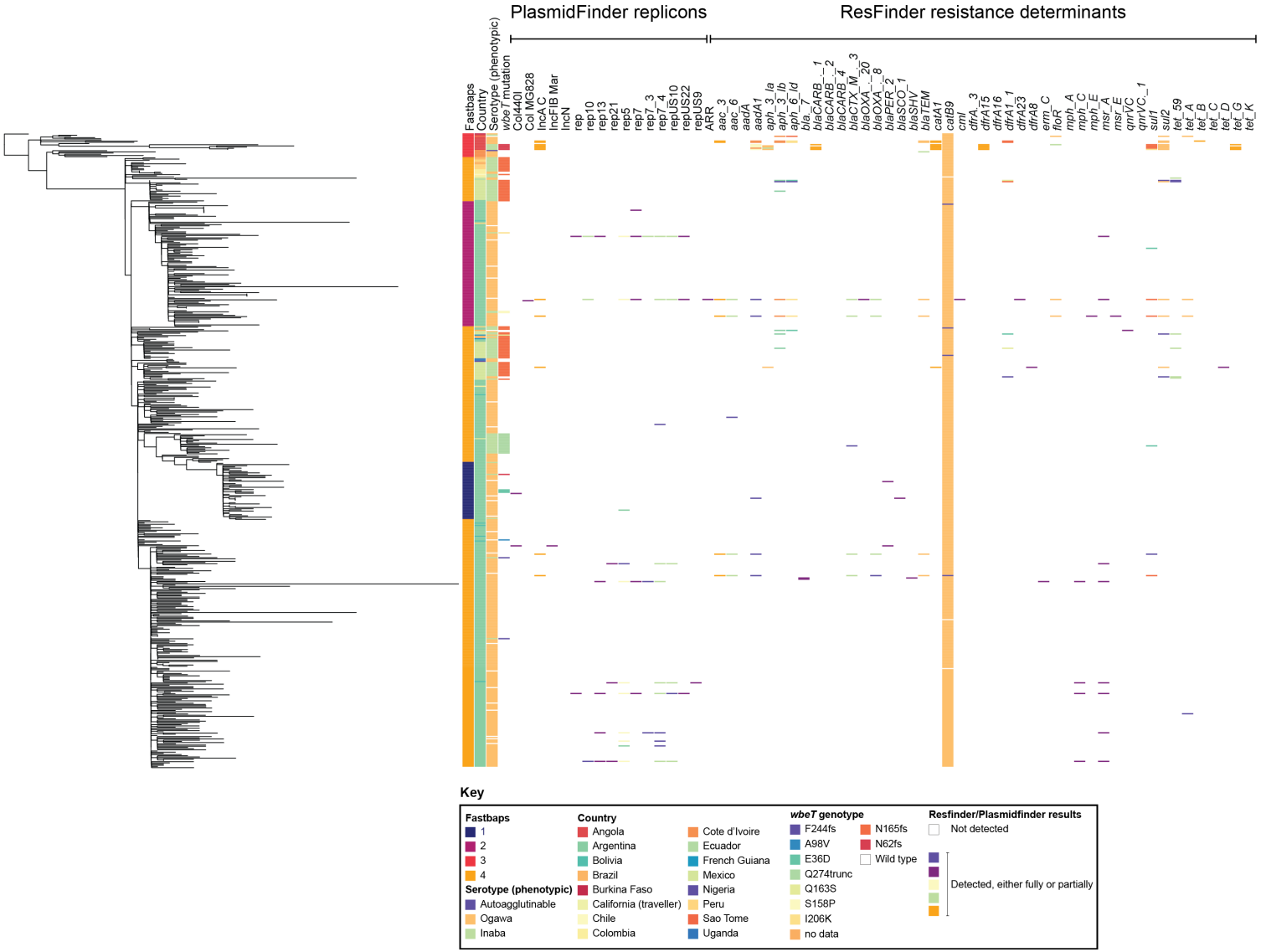

Figure 3.18 – Antimicrobial resistance genes, plasmid replicons, and *wbeT* genotype variants within LAT-1. Data generated using ARIBA (Methods, section 2.1.13).

3.4.6 – Inaba and Ogawa serotype variation within LAT-1

The serotype of *V. cholerae* O1 is of particular historical importance in Latin America. It is known that the initial 1991 cholera epidemics in Peru and elsewhere in Latin America were associated with serotype Inaba *V. cholerae*, which became dominated by serotype Ogawa bacteria in 1992 and thereafter [208, 369]. The Argentinian epidemic began in 1992 and was reportedly dominated by *V. cholerae* Ogawa [387], despite the initial cholera epidemic in Peru being ascribed to *V. cholerae* Inaba [369]. INEI records support this, and also suggest that the dominant serotype in Argentina varied between Inaba and Ogawa over time (Figure 3.3). This suggested subtleties to the dynamics of Ogawa and Inaba serotype *V. cholerae* O1 in Argentina during the epidemic, and begged the question of whether outbreaks due to Inaba bacteria in Argentina after initial Ogawa outbreaks was due to an Inaba strain being imported from elsewhere, or to seroconversion within the lineage already present in Argentina.

As described in the Introduction (section 1.3.1.3), methylation of the terminal sugar on the O1 lipopolysaccharide chain by the WbeT enzyme confers an Ogawa serotype on *V. cholerae* O1, and if the *wbeT* gene is disrupted, WbeT activity is abolished, and an Inaba phenotype results [197–199]. Therefore, to determine a genetic explanation for the apparent flux in serotype in Argentina over the course of the epidemic (Figure 3.3), the genotype of *wbeT* was determined for every isolate in LAT-1. Within the sub-lineage, nine distinct *wbeT* mutations were identified which were predicted to disrupt the WbeT protein (Figure 3.19). It was assumed that mutations that were predicted to frameshift or truncate translated *wbeT* would cause an Inaba phenotype (N62fs, N165fs, F244fs, Q274trunc), as would other mutations that either were found in Inaba isolates in this dataset (I206K) or were otherwise known to confer an Inaba phenotype (S158P [200]). Since all of the isolates harbouring the E36D *wbeT* mutation had an Ogawa phenotype, it was assumed that this mutation does not abolish WbeT function. Of the remaining eight mutations, the genomic predictions correlated well with the phenotypic serotype assigned to each isolate; the *wbeT* genotype matched the phenotypic serotype for all but two of the of the 398 serotyped LAT-1 isolates sequenced in this study (99.4% concordance; Figure 3.19).

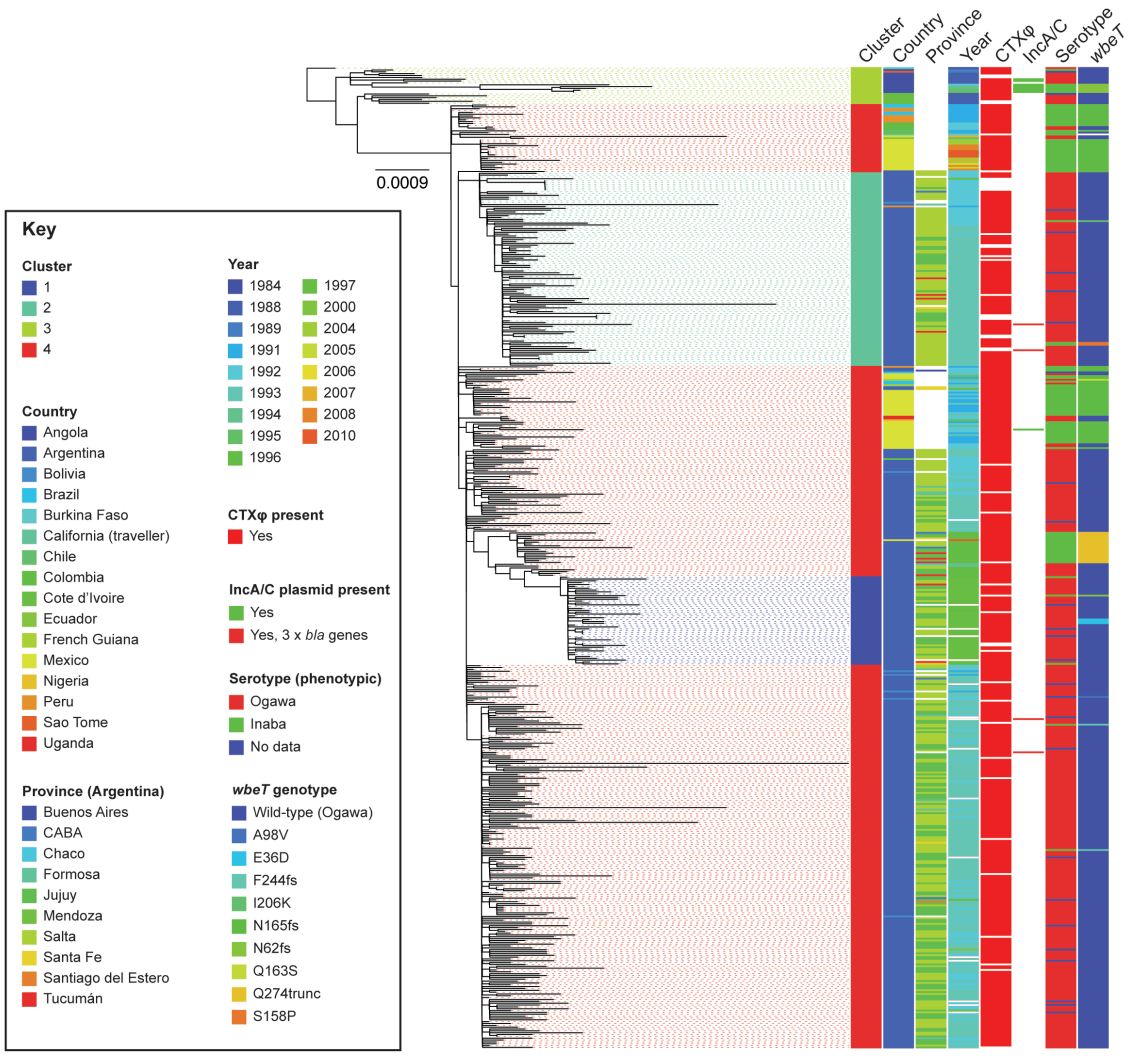

Figure 3.19 – Variation in *wbeT* genotype across the LAT-1 phylogeny. The maximum-likelihood phylogeny presented in Figure 3.14A is re-drawn here with additional metadata and results of select ARIBA analysis results. Missing data are indicated by white space.

Genome data show that the Peruvian Inaba isolates from 1991 harbour the N165fs mutation in *wbeT* (Figure 3.19). Since LAT-1 was introduced into Latin America from West Africa [189], we compared these data to West African Inaba isolates sharing a common ancestor with LAT-1, but post-dating the Peruvian outbreak. These Angolan isolates were collected between 1992 and 1995, just after LAT-1 had been introduced into Peru, and were found to harbour a different mutation, N62fs (Figure 3.19) [158, 189]. Sixty-eight LAT-1 isolates collected since 1991 harbour the N165fs mutation, including isolates from Brazil, Mexico, Chile, Argentina and Colombia, as well as isolates from Peru, all of which were phenotypically serotype Inaba (Figure 3.19). Environmental isolates from Mexico collected between 2004 and 2010 also harbour this mutation, and are part of the same cluster of isolates (Figure 3.19). Hence, the

N62fs and N165fs mutations are likely to have arisen independently, prior to spreading within West Africa and Latin America, respectively.

It has been hypothesised that cholera entered Argentina through the North of the country, which shares borders with Chile, Bolivia, Paraguay, and Brazil [174]. The LAT-1 phylogeny includes genomes from bacteria collected in 1991 and 1992 from Chile, Bolivia and Brazil [189, 234] (Figure 3.19). These were either serotype Ogawa (Bolivia, n = 7; Brazil, n = 1) or Inaba (N165fs; Brazil, n = 6; Chile, n = 1), and were interspersed amongst contemporaneous serotype Ogawa isolates which were collected in Northern provinces of Argentina (Figure 3.19). All were members of cluster 4, except for one Bolivian genome (1992) which was a member of cluster 2, as discussed previously (section 3.4.5; Figure 3.19). Cluster 4 also contains the A1552 reference sequence, which is of an Inaba genotype (N165fs; Figure 3.19). This observation further supports the hypothesis that the same *V. cholerae* sub-lineage circulated within, and between, countries at the Northern border of Argentina.

Once the concordance between phenotypic serotype and genotypic inference had been established, these data were compared to the longitudinal data detailed in Figure 3.3. Following a relative lull in cholera cases in 1995, cases of cholera resurged in Argentina during 1996 [369]. This was associated with serotype Inaba *V. cholerae* (Figure 3.3). Seventeen Argentinian Inaba *V. cholerae* isolates from 1996 formed a closely-related subclade within cluster 4 of the LAT-1 phylogeny (Figure 3.19). These isolates contain a unique mutation in *wbeT*, Q274trunc, and the subclade includes one 2010 Inaba isolate from Mexico (Figure 3.19). Additionally, this subclade shares a common ancestor with the clade of 48 isolates from 1997, which are serotype Ogawa and comprise cluster 1 (Figure 3.19). The 1996/1997 outbreak was not geographically-restricted; this cluster contained isolates from multiple provinces (Figure 3.19).

3.4.7 – Plasmids and antimicrobial resistance in LAT-1

The identification of 3,368 core genes suggested that ~89% of the 3,776 annotated genes in the A1552 reference genome are core to LAT-1. This also implied that gene gain and loss within LAT-1 was also very rare (Figure 3.16; 3.17). However, some examples of gene gain/loss were identifiable; fifty-one of the genomes in the LAT-1 phylogeny were found to lack the CTXφ prophage in its entirety (Figure 3.16; Figure 3.19). It is possible that this loss was a result of long-term culture as has been noted previously [235, 392]. There was also evidence of sporadic

gene gains. For instance, four Argentinian LAT-1 *V. cholerae* were found to harbour genes encoding the extended-spectrum β-lactamases (ESBLs) $bla_{\text{CTX-M-3}}$, $bla_{\text{OXA-8}}$, and bla_{TEM} (Figure 3.18). Manual interrogation of the assemblies for these isolates confirmed that these three ESBL genes were carried on contigs that also included IncA/C plasmid replicons (Figure 3.20).

IncA/C replicons were also detected in Angolan isolates from 1988 and the early 1990s, consistent with previous reports [158, 189] (Figure 3.18, 3.19). These isolates form part of cluster 3 in this dataset, which includes other African genomes pre-dating the transfer of LAT-1 into Latin America in 1991 (Figure 3.18, 3.19). Although two of the Angolan genomes did harbour bla_{TEM} (Figure 3.18), the complement of resistance determinants in these isolates does not match those found in Argentinian *V. cholerae*.

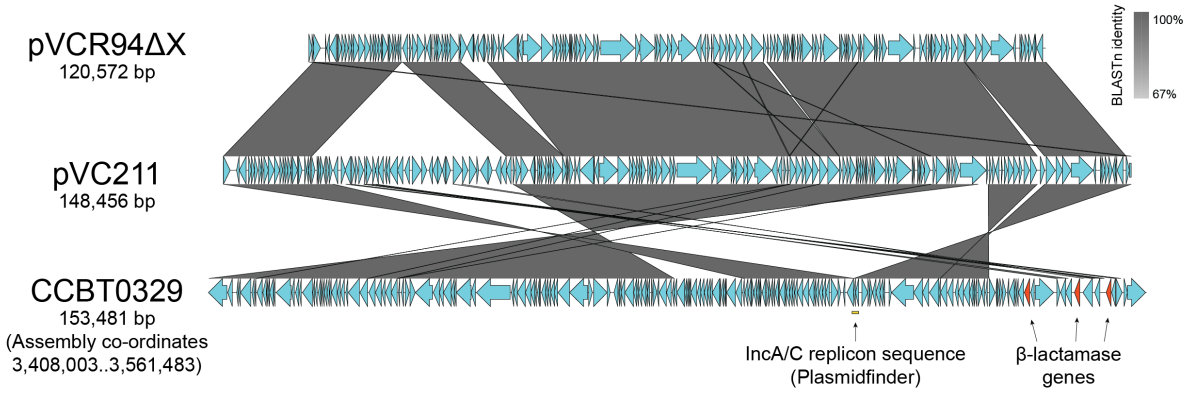

Figure 3.20 – Comparison of a fully-assembled IncA/C2 plasmid from Argentinian isolate CCBT0329 to published *V. cholerae* multidrug resistance plasmids. The positions of the IncA/C replicon sequence present in the PlasmidFinder database and the three genes encoding β-lactamases in the CCBT0329 contig are indicated. The sequences and annotations for pVCR94ΔX and pVC211 were obtained from Genbank (accession # KY399978.1 and KF551948.1, respectively; [393, 394]). The CCBT0329 sequence was annotated with Prokka.

Multidrug resistance plasmids encoding ESBLs have been previously reported in Argentinian *V. cholerae* [366, 387], where molecular characterisations of these large plasmids demonstrated that they can be mobilised and conjugated into *E. coli* [366]. Since IncA/C plasmids are conjugative [395], the plasmids identified in these sequencing data encoding $bla_{\text{CTX-M-3}}$, $bla_{\text{OXA-8}}$, and bla_{TEM} are very likely to correspond to the MDR plasmids seen in *V. cholerae* O1 isolated during the 1990s epidemic from Argentina [366].

3.4.8 – Phylogenetic contextualisation of Argentinian non-7PET isolates

Having performed an in-depth characterisation of LAT-1, and found very limited genetic variation at the levels of SNVs, gene gain and loss, and homologous recombination, attention was turned to the 65 non-7PET Argentinian genomes, to compare the observable dynamics of these isolates to those of LAT-1. In order to contextualise the 65 non-7PET isolates sequenced in this study, these were combined with 318 published non-7PET genome sequences, including three *Vibrio* spp. A pangenome was calculated from these sequences, and 201,790 SNVs extracted from an alignment of 2,719 core gene sequences was used to calculate a maximum-likelihood phylogeny for these isolates (Figure 3.21). The collection of genomes used for contextualisation included a set of recently-reported Chinese genomes [396]; these include isolates which were shown to be closely-related to lineages of *V. cholerae* O1 that had reported to be local to Latin America [189].

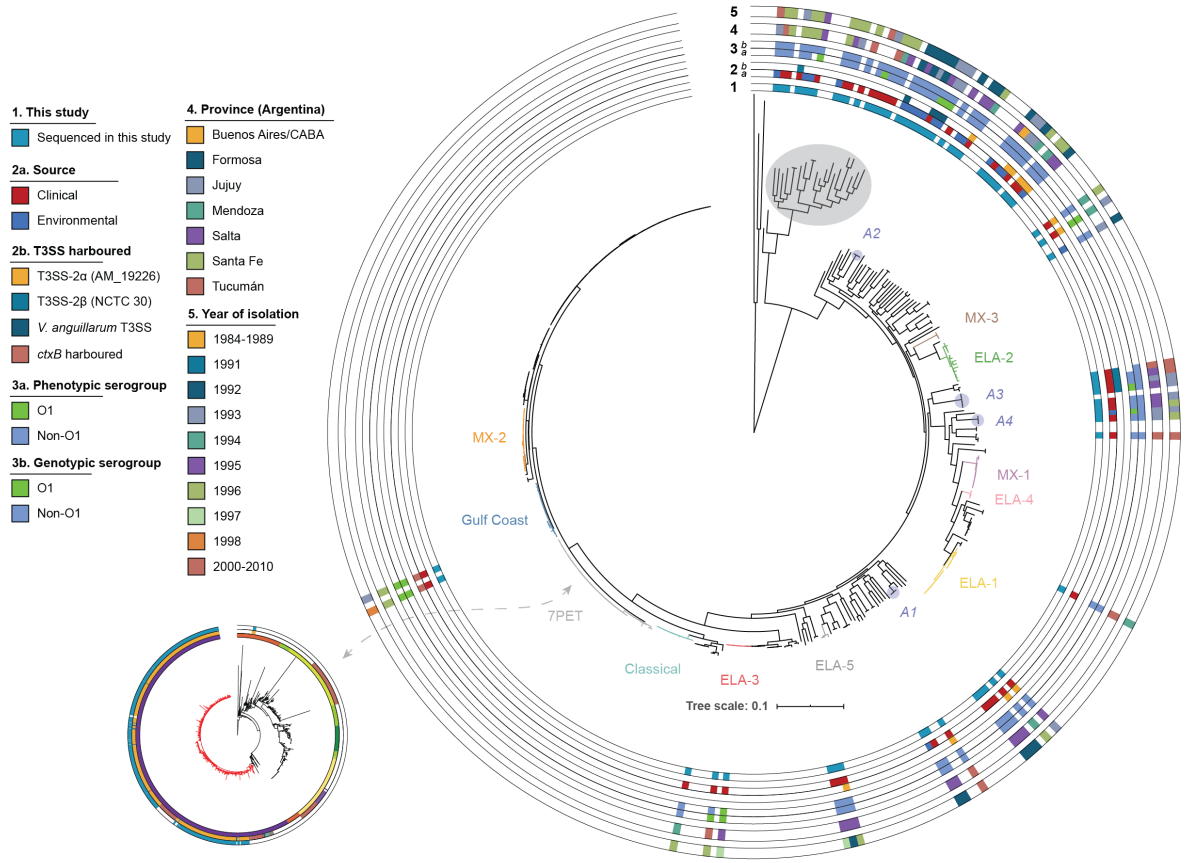

Figure 3.21 – Non-7PET *V. cholerae* phylogeny. A maximum-likelihood phylogenetic tree calculated from 201,790 SNVs identified from an alignment of 2,719 core genes (see Methods). Metadata are presented for the isolates sequenced in this study. The grey disc indicates a cluster of *V. cholerae* which are genetically distinct from pandemic lineages but are still members of the *V. cholerae* species, and will be discussed in greater detail in subsequent chapters. Included in this cluster is a non-toxigenic *V. cholerae* O139 isolate, described in Chapter 4.

In order to confirm the O1 serogroup status of these 65 isolates, the presence of the nucleotide sequence of the serogroup O1 operon extracted from N16961 was determined for each isolate using BLASTn. Four isolates were phenotypically and genotypically serogroup O1. Two of these were members of the previously-described Gulf Coast lineage of *V. cholerae* O1, including the single sequenced *V. cholerae* O1 from 1998 (Figure 3.21). Both Gulf Coast isolates harboured CTXφ and were toxigenic, and the two remaining *V. cholerae* O1 isolates were members of ELA-3 [189] (Figure 3.21). All four isolates were of clinical origin. The remaining 61 isolates lacked the genes required to produce cholera toxin, and were confirmed *in silico* not to harbour genes encoding the O1 antigen, though 45 of these were of clinical origin (Figure 3.21).

Four new lineages of non-O1 non-7PET *V. cholerae* were identified amongst these isolates, defined as clades formed by three or more Argentinian non-7PET isolates in the phylogeny. These were labelled as A1-A4, where A stands for 'Argentina' (Figure 3.21). These lineages contained isolates that were of clinical origin alone (A1, A3) or clinical and environmental origin (A2, A4), were acquired in different years (A3, A4), and from different regions (A2, A3, A4), suggesting that these represent populations of non-7PET *V. cholerae* local to Argentina (Figure 3.21).

3.4.9 – Type III secretion systems in Argentinian non-7PET isolates

Of the 61 non-O1 non-7PET isolates, 21 harboured one of three distinct Type III secretion systems (Figures 3.21, 3.22). These virulence determinants have been associated with pathogenic non-O1 *V. cholerae* as well as with other pathogenic Vibrios [303]. The three T3SSs detected in these isolates included the T3SS-2α described in *V. cholerae* AM_19226, an isolate commonly-used for functional and regulatory studies of T3SS [266, 397]. A less-common system described by Carpenter *et al*., T3SS-2β, was also identified [397], as was a third putative T3SS element which most closely resembles a T3SS detected in two virulent Chilean *Vibrio anguillarum* isolate genomes [398] (Figure 3.22). This putative T3SS was found in lineage A2. The presence of T3SS-2β in lineage A3 was of particular interest – A3 is composed of clinical isolates, contains a previously-described Argentinian isolate, TUC_T2734 [189], and includes one isolate from Salta province collected in the year 2000. T3SS elements were mutually exclusive; no more than one T3SS was detected in a genome. It is also clear from these limited data that more T3SS-positive isolates were of clinical origin

than environmental (T3SS-2α: 10 clinical, 2 environmental; T3SS-2β: 5 clinical, 0 environmental; *V. anguillarum* element; 1 clinical, 3 environmental). No T3SS were detected in 7PET.

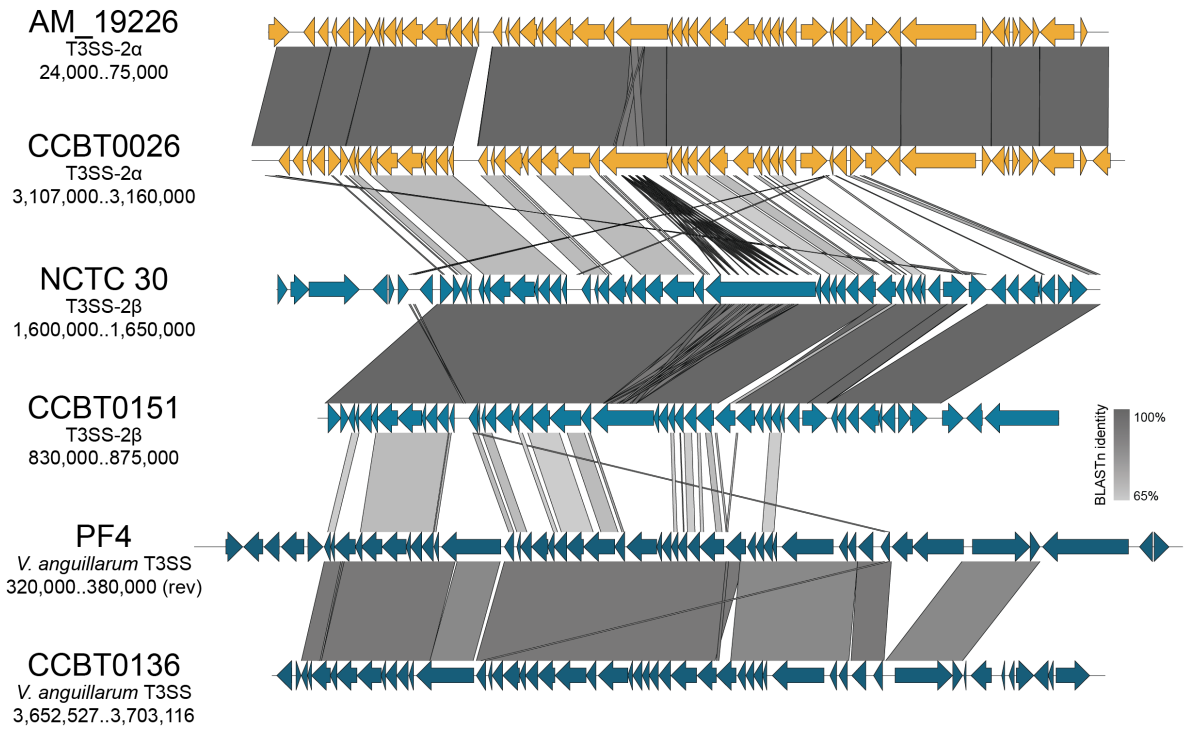

Figure 3.22 – Comparison of T3SS elements detected in Argentinian *V. cholerae* against T3SS taken from reference sequences. Co-ordinates for each region of the assemblies that have been aligned are reported (rev = reverse orientation). Annotations were obtained from Prokka-annotated assemblies used to calculate the pangenome, or from Genbank (PF4 and NCTC 30; accession # CP010081.1 and LS997867.1, respectively). The presence of T3SS-2β in NCTC 30 is discussed in Chapter 4 (section 4.3.10). The PF4 sequence was reflected manually to produce this figure (Adobe Illustrator).

3.4.10 – Comparison of Argentinian LAT-1 and non-7PET pangenomes

Having assigned all of the 490 Argentinian genomes to 7PET or non-7PET, maps of isolate origins stratified on lineage were produced (Figure 3.23). Although non-7PET isolates were fewer in number than 7PET, they were obtained from the same geographical regions and the same time periods as the 7PET isolates (Figure 3.23). This acted as reassurance that the O1 and non-O1 *V. cholerae* sequenced in this study had sampled the same times and regions as one another, albeit that *V. cholerae* O1 had been sequenced in larger numbers than non-O1 (Figure 3.5). This also means that, in principle, both contemporaneous 7PET and non-7PET bacteria should have opportunity to access similar gene pools and to participate in HGT.

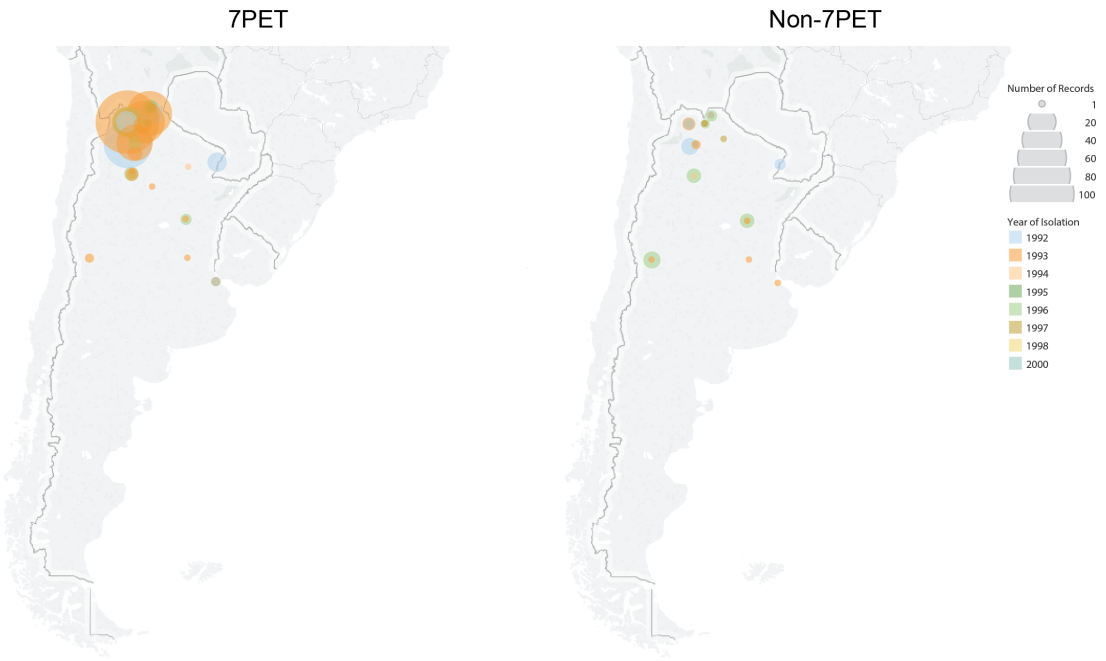

Figure 3.23 – Map showing geographic origin for isolates used in this analysis. Data from Figure 3.4 were stratified by lineage. The size of each circle scales with the number of isolates assigned to that year. Circles coloured by year.

The non-7PET isolates had a considerably expanded accessory genome when compared to LAT-1 (23,458 cloud genes in the collection of 383 diverse genomes compared to 3,313 in the 532 LAT-1 genomes) (Figure 3.24). Thus, the ratio of core genes to cloud genes in LAT-1 was approximately 1:1, whereas in the non-7PET pangenome, the core:cloud gene ratio was ~1:8.

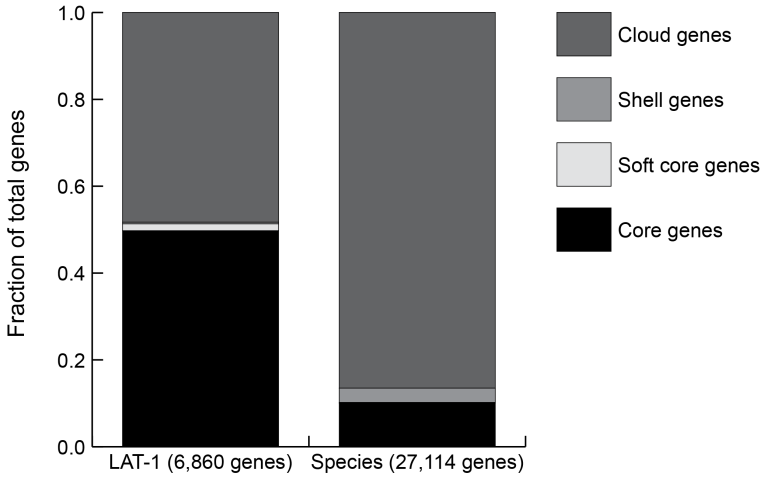

Figure 3.24 – Comparing the summary statistics for the LAT-1 and *V. cholerae* species pangenomes calculated in this study. Fractions of the total number of genes in the pangenome are presented. Definitions used: core: 97% <= strains <= 100%; soft core: 95% <= strains < 97%; shell: 15% <= strains < 95%; cloud: 0% <= strains < 15%.

The large non-core genome was also evident when the pangenome gene presence/absence matrix was visualised against the *V. cholerae* species phylogeny (Figure 3.25). This contrasted starkly with that of the LAT-1 dataset (Figure 3.16).

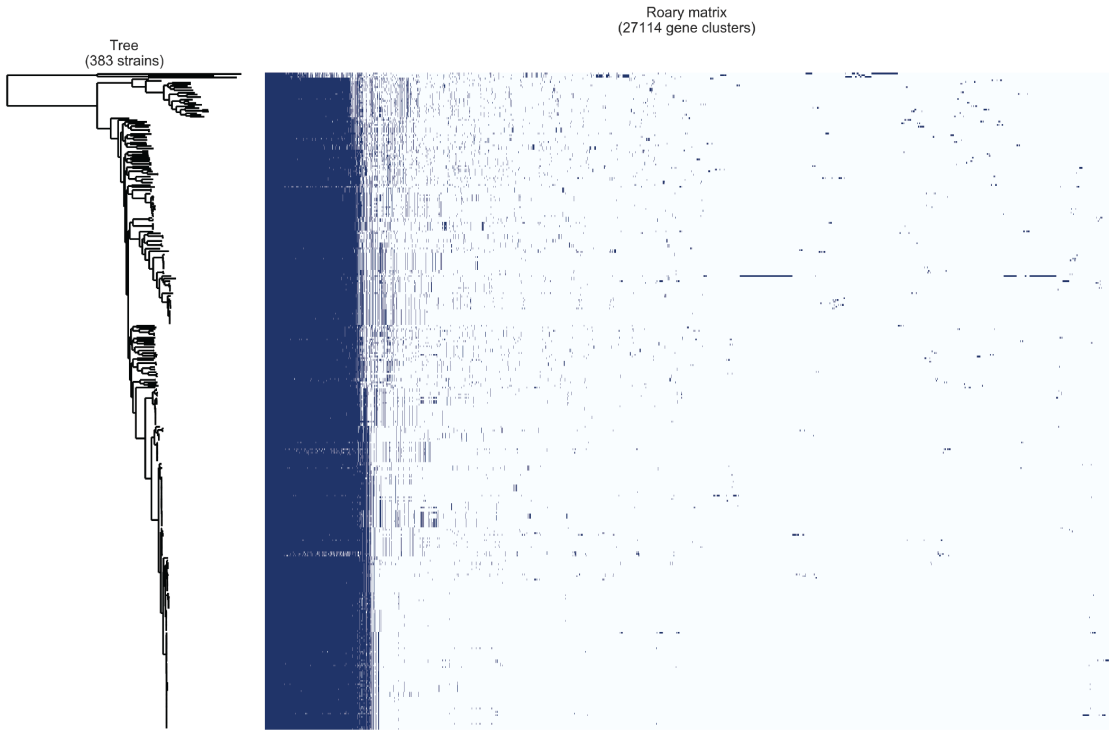

Figure 3.25 – Visualisation of the *V. cholerae* pangenome gene presence/absence matrix. Phylogeny is as presented in Figure 3.21 and is similarly rooted on three *Vibrio* spp. genomes.

The rate of gene discovery as sequences were added to the non-7PET pangenome (Figure 3.26) was much greater than was observed in the LAT-1 pangenome (Figure 3.17), despite there being 38% more sequences in the LAT-1 pangenome. Unlike the LAT-1 pangenome, in which the discovery of new genes was rare, the core of 2,719 genes in this dataset is dwarfed by the number of unique genes discovered as diverse genomes are added to the pangenome. Where the addition of 532 genomes to the LAT-1 pangenome saw ~1,800 unique genes being identified (Figure 3.17), after adding 383 non-7PET genomes to the *V. cholerae* pangenome, ~13,000 unique genes were detected (Figure 3.26).

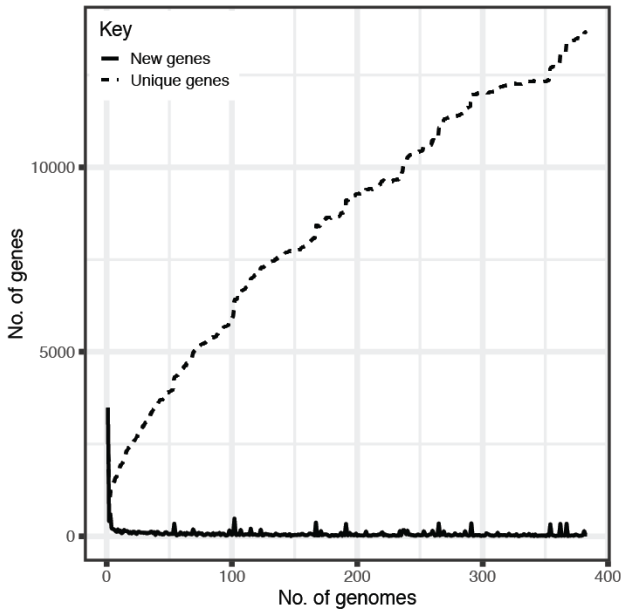

Figure 3.26 – Identification of new genes as genomes are added to the *V. cholerae* pangenome.

The contrast between the clonality of LAT-1 and the diversity of the *V. cholerae* species dataset was also evident when other summary statistics were plotted for these pangenomes (Figure 3.27). Of particular note are the differences in the numbers of total genes relative to conserved genes, the overall contribution to the number of pangenome genes made by adding each additional genome to both datasets, and the number of genes that have a diminishing BLASTp identity in the species pangenome compared to that of LAT-1 (Figure 3.27).

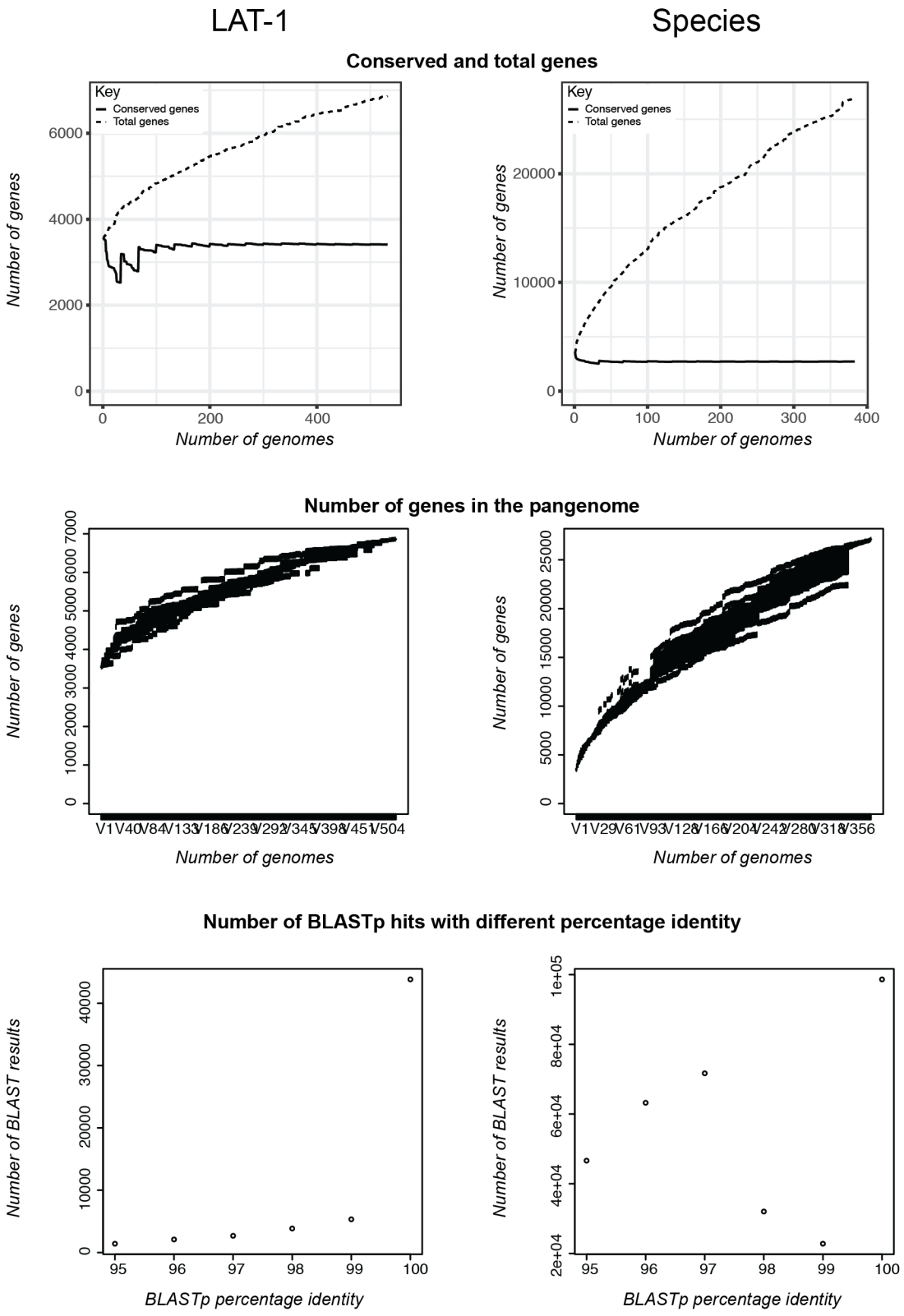

Figure 3.27 – Summary statistics for LAT-1 and non-7PET *V. cholerae* pangenomes.

As another measure of genetic distance, average nucleotide identity (ANI) values relative to the A1552 reference sequence were calculated for all genomes sequenced in this study. By this measure, the non-7PET isolates were also highly genetically diverse in comparison to the 7PET genomes, with a mean ANI relative to A1552 of 97.61 (min 95.90, max 99.65, stdev 0.960; Figure 3.28), in contrast to LAT-1 (mean ANI 99.99, min 99.96, max 99.998, stdev 0.0032; Figure 3.28). This finding is consistent with the phylogenetic analysis presented above (Figure 3.21). It should be noted that an ANI value of 95% is commonly accepted to be threshold for separating species [340].

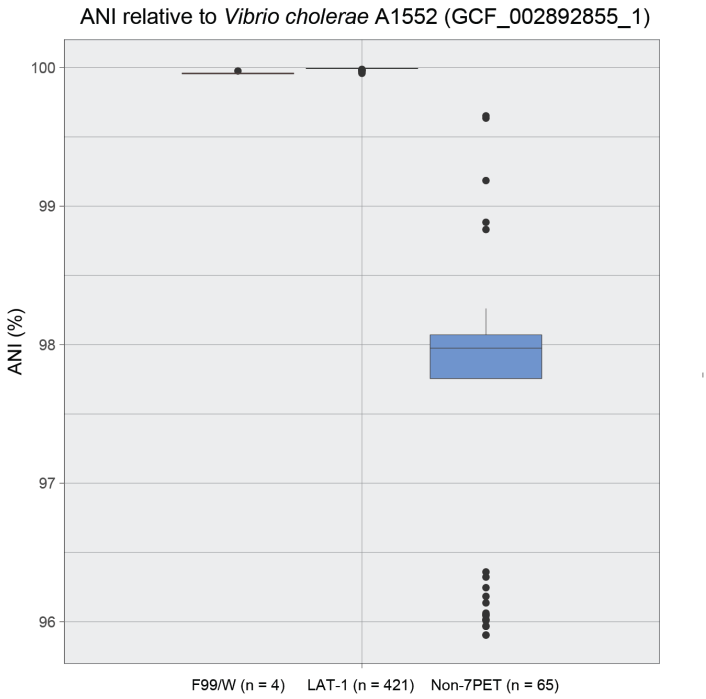

Figure 3.28 – ANI values for genomes sequenced in this study relative to the A1552 reference sequence. Samples were stratified by phylogenetic assignments.

3.5 – Discussion

A combination of factors mean that this study could only have been performed in Argentina. The limited introductions of 7PET sub-lineages into Latin America, the sole introduction of LAT-1 into Argentina, and the enhanced surveillance systems that were introduced in Argentina once cholera broke out in Peru all meant that the INEI culture collection was a comprehensive representation of the 1992-1998 Argentinian cholera epidemic in its entirety. Therefore, the number of isolates included in this project encompassed a very large proportion of the total cholera cases across the country and throughout the epidemic, presenting a unique opportunity to understand the dynamics of epidemic *V. cholerae* evolving over long periods, originating from point-source introductions. To the best of our knowledge, this is the largest genomic study to date that investigates a cholera epidemic in a single country. We also believe that it is the largest genomic analysis of any single bacterial pathogen in Argentina.

There are a number of observations to be made using these data, chief among which is the fact that it is clear that a single clone of *V. cholerae* O1, now known to be one sub-lineage of 7PET [189], was responsible for pandemic cholera in Argentina, in spite of the seasonal fluctuations and serotype variation observed (Figure 3.3) [274, 367, 399, 400]. These data also show that LAT-1 circulated amongst the countries at the Northern borders of Argentina during the early 1990s - for instance, cholera was first reported in Bolivia in August 1991 [401, 402], and Bolivian genomes from the early 1990s are mixed amongst the Argentinian genomes from the same period (Figures 3.14, 3.19). These genomic data also validate fundamental observations made by public health authorities during the cholera epidemics of the 1990s, such as that Argentinian cholera outbreaks were principally caused by *V. cholerae* O1 serotype Ogawa which had been shown by PFGE to be closely related to the Peruvian strain [208, 367, 403–405].

These data also explain historical observations on the variation of serotype during the 1990s epidemic across Latin America, and provide new insight into the microevolution of LAT-1 during the Argentinian cholera epidemic. *V. cholerae* Ogawa from Argentina in 1992 were shown to be closely related to Inaba isolates from Peru (Figure 3.19), and the shift in dominant serotype from Inaba to Ogawa that was observed in Peru and elsewhere in Latin America [208] represented variation within LAT-1, rather than a separate introduction of another strain (Figure 3.14A). Similarly, the outbreak of Inaba *V. cholerae* in Argentina in 1996 was also a

result of variation within LAT-1; the outbreak was caused by LAT-1 isolates in which *wbeT* had mutated from wild-type Ogawa genotype to an Inaba genotype (Q274trunc). This mutation is likely to have occurred in Argentina. The Q274trunc mutation is distinct from others identified within LAT-1, particularly the mutation associated with the Inaba phenotype in contemporaneous Mexican isolates (N165fs). These results strongly indicate that Argentinian cholera in 1996 was not caused by an introduction of a new Inaba (sub)lineage from elsewhere in Latin America; rather, LAT-1 *V. cholerae* that had already been introduced into Northern Argentina, or neighbouring countries, acquired a new Inaba genotype. In turn, the Argentinian cholera outbreak in 1997 was caused by a sub-clade of LAT-1 that was closely related to the 1996 Inaba clone. However, the topology of our phylogeny suggests that this was not a result of 'reversion' from the Inaba Q274trunc genotype to an Ogawa genotype (Figure 3.19). These data underline that Ogawa/Inaba phenotypic variation is not phylogenetically informative, and may be both misleading and inappropriate to use as an epidemiological marker.

Furthermore, the inclusion of non-7PET *V. cholerae* in our study has highlighted that a highly diverse population of the *V. cholerae* species existed in Argentina concurrently with the extremely invariant LAT-1 pandemic sub-lineage during the 1990s. It is suggested that these non-7PET bacteria, including serogroup O1 and non-O1 isolates, represent those *V. cholerae* that are truly endemic to Argentina, which are evolving locally, but lack the propensity or ability to cause global epidemics and to spread in the same way as 7PET. Therefore, it can be concluded that the reason that Latin America was cholera-free for 97 years was due solely to the absence of pandemic *V. cholerae* lineages from the continent. In the absence of comprehensive clinical data associated with these non-7PET isolates, it cannot be determined whether they are aetiological agents of cholera, or of a cholera-like illness. However, it is very clear that non-7PET *V. cholerae* were present in Argentina, and associated with disease at a low level, throughout the 1992-1998 cholera epidemic and thereafter (Figure 3.3).

In spite of the sustained circulation and dissemination of LAT-1 across Northern Argentina, an area of approximately 1.2 million km^2 (Figure 3.4), these data suggest that very little genetic change, in terms of SNVs, recombination, and gene gain/loss, occurred in this sub-lineage over a period of nearly six years. The invariance of LAT-1 is juxtaposed with the diversity observed in non-O1 *V. cholerae* in Argentina (Figures 3.14A, 3.16, 3.21, 3.24-3.28). As discussed in the Introduction to this chapter, although *V. cholerae* research has focused on studying epidemics and outbreaks, by definition, this tends to describe epidemic lineages. Non-7PET *V. cholerae*

are highly variable – within this dataset, as well as examples of local lineages of non-7PET *V. cholerae*, isolates were also identified which were confirmed microbiologically to be *Vibrio cholerae*, but were diverse phylogenetically (Figure 3.21) and as measured by ANI values (Figure 3.28). The pathology of the disease associated with these isolates – and whether virulence determinants such as T3SS contribute to this disease – is beyond the scope of this thesis but is the focus of future work, though evidence does suggests that T3SS do contribute to diarrhoea and disease caused by non-7PET bacteria [305]. With the caveat of a small sample size, it can also be observed that the clinical non-7PET isolates were enriched for the presence of T3SS (16/21 isolates).

It is particularly vital to understand the diversity of local, endemic *V. cholerae* that co-exist alongside 7PET during a cholera epidemic, because non-epidemic *V. cholerae* present in a country may contribute to disease that is symptomatic of cholera, but does not pose the same relative risk to public health as 7PET [189]. Similar observations have recently been made in China [396]. The relative risk of *V. cholerae* lineages should be accounted for in the magnitude of epidemic preparedness responses to such outbreaks. This is particularly relevant in the wake of the GTFCC commitment to reducing deaths from cholera by 90% before the year 2030 [9]. This campaign focuses on the control of cholera, the disease, rather than on controlling 7PET, the aetiological agent of pandemic cholera. As cholera control is implemented, countries experiencing a high incidence of cholera attributable to 7PET will see a decline in the number of cholera cases. Therefore, it is anticipated that as pandemic cholera reduces in magnitude, disease caused by non-7PET *V. cholerae* will become more visible, as was observed in Argentina and Latin America. By using genomics to differentiate pandemic and non-pandemic lineages for public health epidemic preparedness responses, concerted control efforts including epidemiologists, public health authorities and microbiology laboratories targeting 7PET specifically, and accounting for background levels of endemic non-7PET disease, could see epidemic cholera eliminated in Latin America.

To summarise, as well as describing the genomic history of cholera in Argentina during the 1990s, this chapter describes the stark contrasts can be observed during a country-wide cholera epidemic between the clonality of 7PET and the diversity of contemporaneously-isolated non-7PET *V. cholerae*. This suggests that to understand more about the disease, particularly the disease that can be caused by non-epidemic and non-pandemic *V. cholerae*, our efforts must

begin to turn from the exclusive study of pandemic *V. cholerae* towards a more holistic and concurrent analysis of pandemic and non-pandemic isolates.

This chapter has focused on analysing LAT-1 in detail, and has made high-level comparisons between the LAT-1 and species datasets. Clearly, although non-7PET *V. cholerae* continue to be associated with clinical cases of disease, they remain understudied. In Chapter 4, a detailed study of a small number of historically- and medically-important non-7PET non-O1 *V. cholerae* will be described, using both *in silico* and *in vitro* approaches. This will focus on characterising specific aspects of their genomes. Once these isolates have been described in fine-detail, subsequent chapters will be of broader scope, return to studying larger collections of genomes *in silico* with the aim of extrapolating the results of these analyses into the context of larger collections of diverse genomes. The non-7PET genomes described in this chapter will also be re-examined in subsequent chapters, to consider aspects such as antimicrobial resistance, virulence gene distribution, and plasmid replicons, in the context of additional diverse genomes and the data presented in Chapter 4.