Chapter 5

The accessory genome - concordance and conflict between *V. cholerae* genomics and phenotypic dogma

Contribution statement

Nick Thomson supervised the work described in this chapter. NCTC cultures were supplied by Sarah Alexander and Julie Russell (NCTC). Jake Turnbull assisted with the collation of NCTC internal records. Additional isolates were supplied by Florian Marks (IVI) and Claire Jenkins (PHE). The Mexican isolates and metadata described here were supplied by Alejandro Cravioto. Charlotte Tolley extracted gDNA from a limited number of live isolates at WSI under my supervision.

I performed all experiments and analyses, and produced all figures.

COVID-19 statement

The work described in this chapter was affected by the shutdown imposed on the University of Cambridge and the Sanger Institute by the COVID-19 pandemic. This affected planned experiments to validate *in vitro* the genomic observations described in this chapter.

5.1 – Overview

Chapters 3 and 4 have both highlighted the striking differences in diversity between pandemic and non-pandemic *V. cholerae*. Chapter 3 described a focused study of a clonal sub-lineage of pandemic *V. cholerae*, and presented an initial characterisation of 65 non-pandemic *V. cholerae* from Argentina. In Chapter 4, an effort was made to explore the diversity of *V. cholerae* beyond the pandemic lineage by focusing specifically on the forensic analysis of a small number of closed genome sequences, taken from specific *V. cholerae* that were of biological and historical interest.

Much of our understanding of the differences between pandemic and non-pandemic V. cholerae, particularly at the level of gene expression and regulation, come from detailed molecular studies of a handful of reference strains of bacteria. These include 7PET isolates such as N16961, the strain used for the initial *V. cholerae* sequencing study, and C6706 and A1552, both from Latin America [54, 59, 378]. Classical isolates such as O395 [97] and 569B [219, 435] have been well-characterised, as have a handful of serogroup O139 isolates [436] and non-O1 isolates such as V52 and AM_19226 [256, 266]. However, the insights gleaned from these laboratory strains are rarely extrapolated into a wider genomic context, to understand how key regulators of gene expression or phenotypic determinants are distributed and vary across a diverse population of bacteria.

Expanding horizons to consider larger and more diverse collections of genome sequences increases the context into which our understanding can be placed. For example, by studying a single representative isolate of classical biotype *V. cholerae* (O395), it was initially hypothesised that all classical biotype (and thus, Classical lineage) isolates possessed a mutation in *hapR*, preventing HapR-dependent regulation of gene expression in response to cell density [437]. However, considering additional genomes caused this generalisation to be disproven – although O395 does indeed have an inactivating mutation in *hapR*, this is not the case in all Classical isolates, such as CA401[438].

Having access to as many diverse genomes as possible is therefore essential to allow for a maximally-unbiased study of the *V. cholerae* species. For instance, it has been stated previously that in *V. cholerae*, plasmids are rare, and drug resistance is usually encoded by conjugative transposons in the bacterial chromosome [230]. In order to make accurate determinations about

the frequency of plasmid types or antimicrobial resistance determinants across a species, it is essential not to focus exclusively on 7PET – as has been demonstrated in earlier chapters in this thesis, the dynamics of 7PET and non-7PET *V. cholerae* are extremely different.

In this chapter, I present an analysis of a collection of diverse, non-pandemic *V. cholerae*. These were collated from several sources, and include a set of historically-important isolates from the NCTC collections, a set of recent isolates from travellers returning to the United Kingdom [439], and non-O1 isolates of both clinical and environmental origin from Mexico. These were added to the diverse phylogeny of isolates presented in previous chapters, to expand as much as possible the sequenced diversity of *V. cholerae*.

5.2 – Specific aims

In this chapter, I aim to:

1) Determine the population structure of a large collection of genomically-diverse *V. cholerae* isolates,
2) Characterise the distribution of key virulence, antimicrobial resistance genes across the dataset,
3) Characterise the genomes of a number of historically-significant isolates, and use these to understand phenotypic observations made about these isolates in the past, and
4) Explore the molecular basis of biotypes across the *V. cholerae* species.

The isolates described in this chapter were particularly valuable because information on their history and provenance was available for analysis. The dataset consisted of a wide variety of *V. cholerae* collected as part of several different projects. In this chapter, a targeted approach was taken to characterise the distribution across the species of genetic determinants that are specifically associated with important phenotypes.

5.3 – Results

5.3.1 – Expansion of the V. cholerae *species phylogeny*

Two hundred and sixty-eight additional genomes were added to the 383 genomes used to compute the non-7PET phylogenies presented in Chapters 3 and 4. A deliberate effort was made to include non-O1 *V. cholerae* when compiling these genomes, to attempt to maximise the genetic diversity within the dataset. After excluding contaminated and poorly-assembled sequences, a pangenome was calculated using the final collection of 651 sequences (11 isolates were sequenced twice, meaning that 640 independent sequences are represented in this dataset). Using 187,675 variable sites in a core-gene alignment of 2,721 genes, a maximum-likelihood phylogeny for these isolates was calculated. Statistical support for this phylogeny was determined using 5,000 ultrafast bootstrap approximations (Figure 5.1).

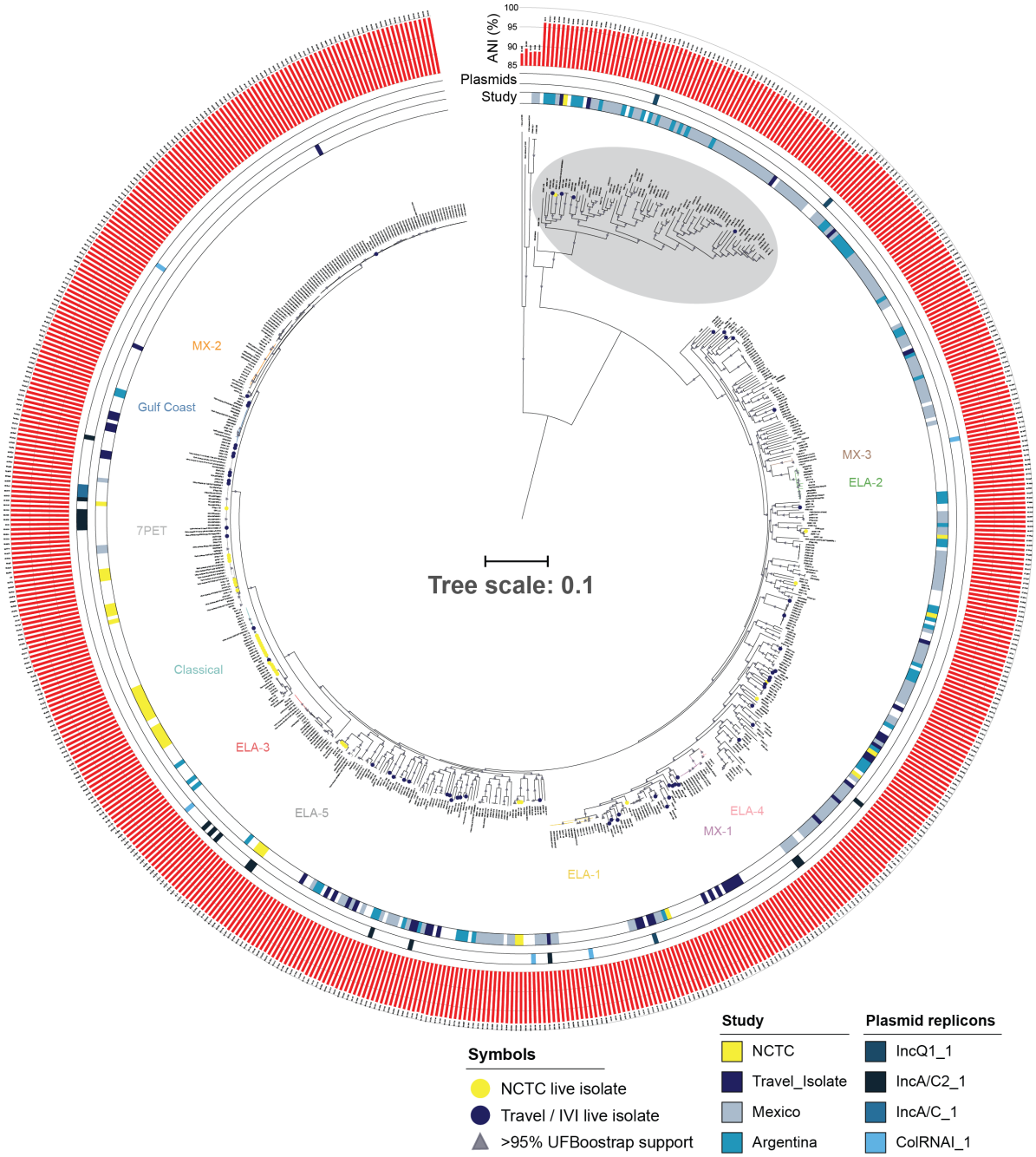


Figure 5.1 – A maximum-likelihood phylogeny of 646 *V. cholerae* and 5 *Vibrio* spp. Triangles on nodes indicate a node that has >95% UFBootstrap support (5,000 replicates). The outer ring denotes the strain collection from which each sequence was obtained. Coloured circles on leaves indicate that a live culture of the sequenced isolate was available at WSI for subsequent experimental analyses. Lineages were coloured and named for consistency amongst thesis chapters. Height of red bars indicates ANI percentage relative to the A1552 reference sequence as described in Chapter 3. Plasmid replicons were identified using ABRicate and the PlasmidFinder database (see below). The tree is rooted on the *Vibrio* spp. outgroup. Scale bar denotes substitutions *per* variable site. The clade of isolates to which NCTC 30, 48853_F01 and multiple Argentinian non-O1/O139 *V. cholerae* belong is highlighted in each case (grey disc).

5.3.2 – Initial characterisation of diverse V. cholerae

It was apparent that two of the Mexican isolates sequenced in this study clustered adjacent to the *Vibrio* spp. outgroup (Figure 5.1). Inspection of the Kraken QC report for these two isolates confirmed their identity to be *Vibrio metoecus*. These two sequences were retained as part of the outgroup for subsequent analysis. A clade of divergent *V. cholerae*, to which NCTC 30 and the non-toxigenic *V. cholerae* O139 (48853_F01) belong, was also substantially expanded as a result of adding these diverse genomes (Figure 5.1).

Comparing the topology of *V. cholerae* species phylogenies from Chapters 3 and 4 to the phylogeny discussed in this chapter provided an intuitive overview of the diversity being discovered as non-O1 isolates were added to the sequence collection (Figure 5.2).

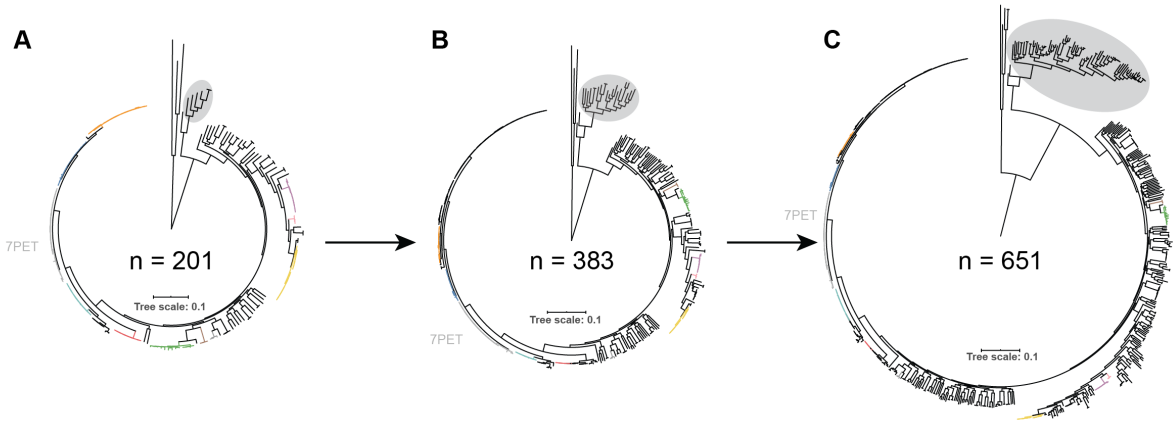


Figure 5.2 – Illustrating the iterative expansion of the *V. cholerae* phylogeny during this thesis research. A: The phylogeny from Figure 4.26, consisting of previously-published isolates and NCTC 30. B: The phylogeny from Figure 3.21, which consists of the dataset used in (A), Argentinian *V. cholerae*, and published Chinese genomes. C: Adding 217 genomes for this chapter to the phylogeny in (B). This phylogeny is presented in detail in Figure 5.1. Lineages are coloured as in Figures 3.21, 4.26 and 5.1. Scale bars denote substitutions *per* variable site. All phylogenies rooted on *Vibrio* spp.

In addition to phylogenetic analysis, average nucleotide identity (ANI) calculations relative to the A1552 reference were performed for each genome assembly, to provide an overview of the relative differences in nucleotide diversity across the phylogenetic tree. These data show that sequences in this divergent clade have a mean ANI of 96.15% relative to A1552 (min 95.85, max 96.42), which is significantly different both to the remaining *V. cholerae* (mean 98.65, min 97.54, max 100) and *Vibrio* spp (mean 88.6, min 88.0985, max 89.3209) (Figure 5.3).

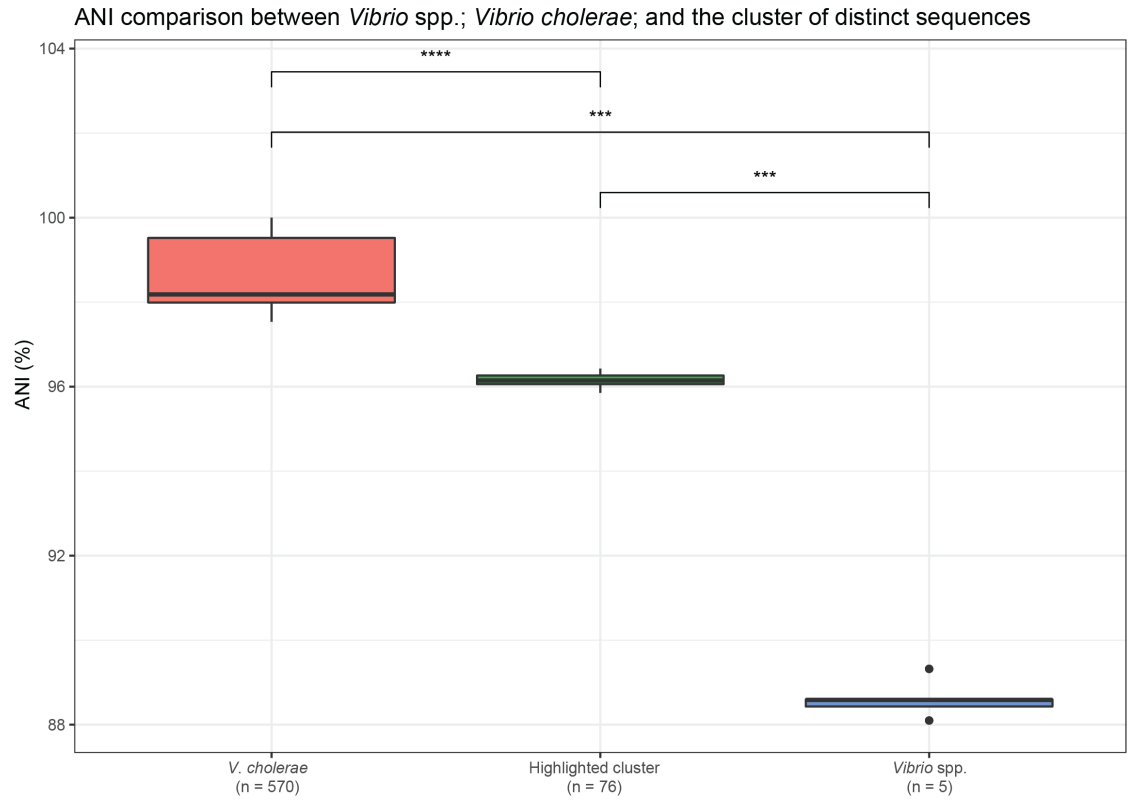


Figure 5.3 – Comparing ANI values for *V. cholerae*, *Vibrio* spp., and the group of diverse sequences highlighted in Figures 5.1 and 5.2. A Wilcoxon rank-sum test was used for statistical testing, to avoid assuming that ANI data were normal in nature. ***: $p < 0.001$, ****: $p < 0.0001$. The highlighted cluster corresponds to isolates contained within the grey disc in Figure 5.1.

The genomic heterogeneity of this divergent clade relative to the remainder of *V. cholerae* in the dataset was also illustrated by the gene presence/absence matrix for this pangenome (Figure 5.4). In this visualisation, clusters of genes absent from this cluster but present in the remainder of sequenced *V. cholerae* are evident.

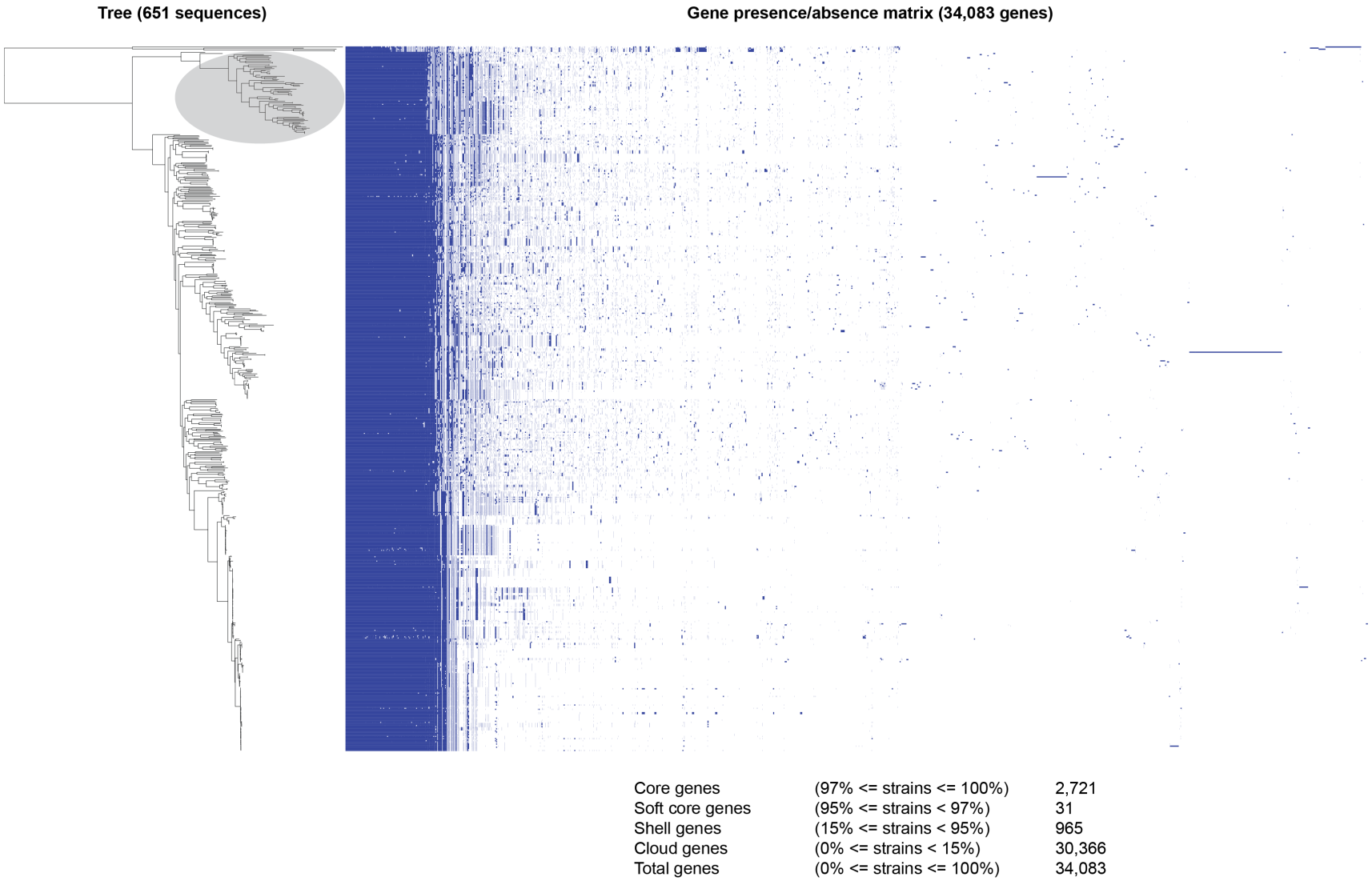
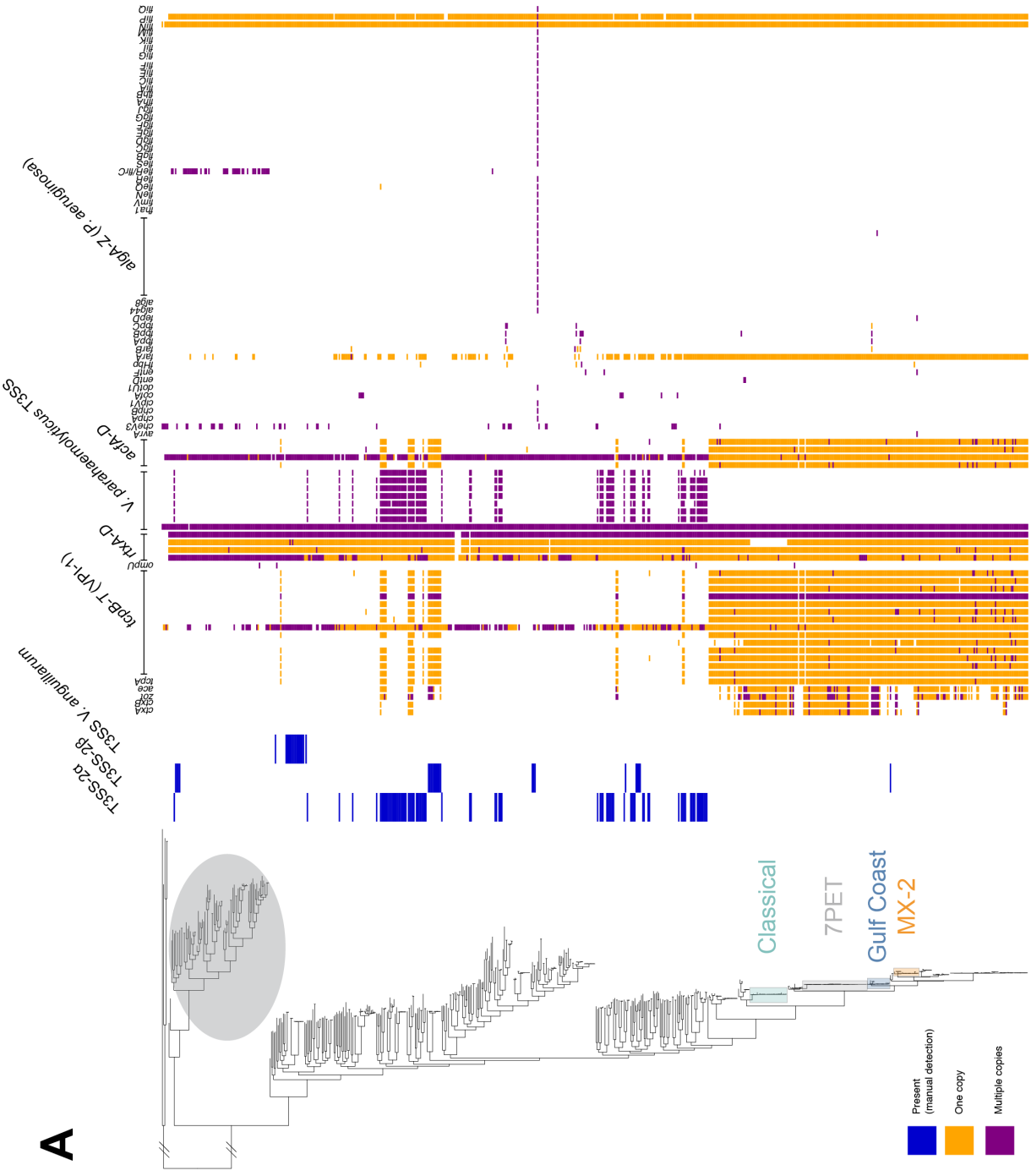


Figure 5.4 – Summary statistics and visualisation of the gene presence/absence matrix for the expanded *V. cholerae* phylogeny. The grey disc indicates the set of diverse *V. cholerae* discussed in section 5.3.2

5.3.3 – Virulence gene distribution across the V. cholerae *phylogeny*

To begin to explore the diversity of these genomes, the distribution of orthologues of known virulence determinants across the dataset was determined. Many of the non-pandemic isolates for which genomes were available were obtained from clinical sources – *i.e.*, they caused an illness in a human patient. Although the clinical metadata for these isolates are sparse, determining the possible mechanisms by which these isolates might have caused illness from their genome sequences was performed. To do this, the presence of genes included in the VFDB database was determined across the pangenome. In addition, the analysis of Argentinian non-7PET genomes in Chapter 3 had identified genes corresponding to three distinct T3SS. These were used to determine whether such systems were present in this expanded dataset using the pangenome gene presence/absence matrix. These results are presented in Figure 5.5.


A
T3SS-2α
T3SS-2β
T3SS V. angullarum
tcpB-T (VPI-1)
tfxA-D
V. parahaemolyticus T3SS
actA-D
aIgA-Z (P. aeruginosa)
Classical
7PET
Gulf Coast
MX-2
Present (manual detection)
One copy
Multiple copies

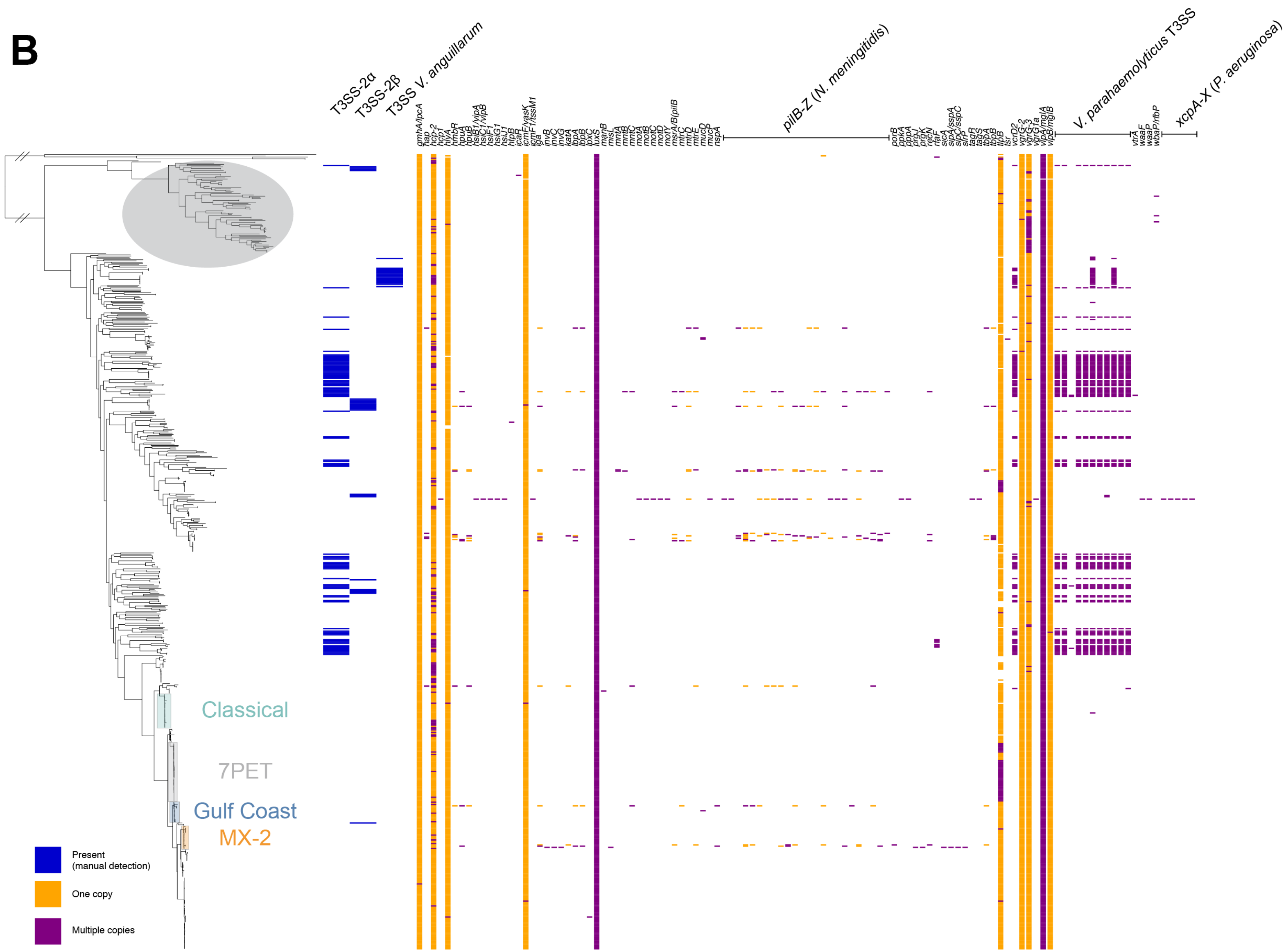


Figure 5.5 – Distribution of key virulence genes across the *V. cholerae* phylogeny. Legend on adjacent page

Figure 5.5 legend: The presence of T3SS elements was determined by identifying genomes containing orthologues of the T3SS genes characterised in Chapter 3. The genome for isolate SRR221551 was the only sequence to contain a large set of *P. aeruginosa* virulence genes. Manual inspection of this assembly and verification of its Kraken report demonstrated that this sequence was contaminated with *Pseudomonas* sequence. Hatch marks denote branches which have been manually shortened for illustrative purposes. Figure has been split into two sub-figures for legibility purposes.

Clusters of T3SS genes appear to be distributed amongst the diverse *V. cholerae* in the dataset, though they are absent from the epidemic lineages and the lineages related to these (Figure 5.5). It also appears that T3SS are rare amongst the clade of highly-diverse isolates, though T3SS-2β was detected in NCTC 30 and related isolates (Figure 5.5). The presence of *V. parahaemolyticus* T3SS genes, as identified by ABRicate using VFDB, appears to indicate the presence of T3SS-2α (ABRicate did not identify complete T3SS in genomes harbouring either the T3SS-2β or the *V. anguillarum* T3SS described in Chapters 3 and 4). However, two of the T3SS-2α genes included in VFDB were detected in isolates harbouring *V. anguillarum* T3SS, suggesting that these genes are conserved amongst or common to the two systems. Genes encoded by VPI-1 were found throughout pandemic lineages, related genomes, and in diverse isolates within the dataset. No isolate harboured *ctxAB* in the absence of VPI-1, consistent with our understanding that VPI-1 encodes the TCP receptor to which CTXφ binds, notwithstanding transduction of CTXφ amongst TCP$^-$ *V. cholerae* [93]. T3SS elements were not mutually exclusive with either CTXφ or VPI-1. This is not the same as the observations made in Chapter 3, and underlines the fact that as more diverse *V. cholerae* are sequenced, our understanding of gene distribution within the species will be refined.

One clade of isolates was identified which harboured the T3SS most similar to a system from *V. anguillarum*, described in Chapter 3 (Figure 5.5). This cluster of isolates was polyphyletic, consisting of one clade of the three Argentinian genomes described in Chapter 3 and a second clade comprising an additional 11 non-O1 isolates from Mexico and Guatemala (eight of clinical origin, three of unknown origin). None of these genomes appear to contain any other known *V. cholerae* virulence determinants. This observation underlines further the fact that there is a great need to understand the contributions made by T3SS to disease caused by *V. cholerae* in humans – throughout this thesis research, several examples have been identified of clinically-isolated *V. cholerae* that lack putative virulence determinants other than T3SS, justifying further research into the roles that such systems may play in causing acute watery diarrhoea or symptoms of cholera.

5.3.4 – Serogroup assignment of isolates in silico

Complete metadata were unavailable for the sequenced isolates contained in this analysis. Therefore, the serogroup (O1 or non-O1) of each isolate was confirmed *in silico*. To do this, the presence and absence of each of the genes in the O1 biosynthesis gene loci was determined for each isolate in the dataset (Figure 5.6A). These loci have been delineated previously in *V. cholerae* (discussed in Chapter 4 in the context of *V. cholerae* O139). Isolates that are known not to be serogroup O1 – including serogroup O139 and O37 isolates – were confirmed to lack genes present in serogroup O1 isolates (Figure 5.6B). Based on these results, all of the isolates in the dataset were determined to be O1 or non-O1 *V. cholerae*.

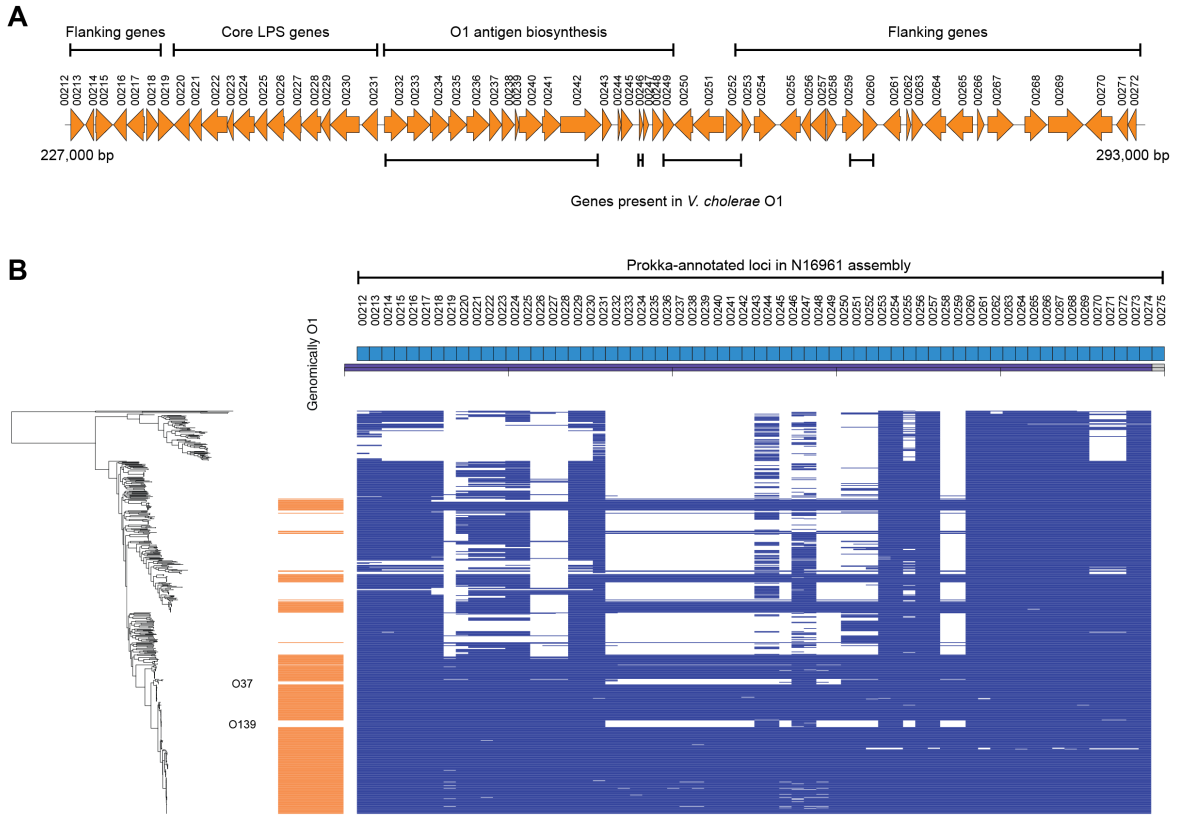


Figure 5.6 – The distribution of O1 antigen biosynthesis genes in the *V. cholerae* pangenome. (A): Schematic of the genomic organisation of the O1 biosynthetic locus in the N16961 reference sequence annotated with Prokka for uniformity within the pangenome dataset. The genes involved in producing the core lipopolysaccharide as well as the O1-specific LPS antigen are denoted. (B): The presence and absence of the 60 genes labelled in (A) across the pangenome of 651 isolates presented in Figure 5.1. Isolates which harbour all of the genes in (A) were determined in silico to be serogroup O1. Where required, this was also validated using BLASTn to search the assembly for the presence of the O1 biosynthesis locus in its entirety, as carried out in Chapter 3. This approach allows for the easy recognition of non-O1 *V. cholerae*; the O37 and epidemic O139 clusters (7PET sub-lineage) are indicated as examples.

5.3.5 – Distribution of key pathogenicity islands amongst V. cholerae

It has been stated repeatedly in the literature that the four major *V. cholerae* pathogenicity islands, VPI-1 and VPI-2, VSP-1 and VSP-2, are markers for epidemic and pandemic lineages of *V. cholerae* O1 [133, 142]. Having determined whether or not each isolate in the phylogeny was serogroup O1, and knowing from Chapter 4 that these elements may be present in more than one copy, the presence/absence and copy number of the genes encoded by these pathogenicity islands across the phylogeny was determined. This was done by testing for the presence of each of the genes that comprise these islands in the pangenome data, in order to obtain a sense of the variability within these elements across the phylogeny. These results are summarised in Figure 5.7.

Some previously-described phenomena were evident amongst these results, such as the VPI-2 deletion in *V. cholerae* O139 [134, 409], and the deletion in VSP-2 reported in recent 7PET isolates [309]. As previously observed, a number of non-7PET *V. cholerae* O1 from China were found to harbour VSP-1 in its entirety [396], as does a non-7PET *V. cholerae* O1 isolate from Thailand [440]. These results are of particular importance because they challenge the hypothesis that VSP-1 and VSP-2 are specific to 7PET [133]. It also appeared that genes orthologous to those encoded by VSP-2 were identifiable in multiple *V. cholerae* that were distantly related to 7PET (Figure 5.7). It was also evident that a number of non-O1 isolates in this dataset harboured VPI-1 (genes on which encode TCP), suggesting at least in principle that these non-O1 isolates could be lysogenised by CTXφ. However, no non-O1 isolates harbouring VPI-1 and CTXφ were detected amongst these data (Figure 5.7).

This analysis underscored the importance of using closed genome assemblies for such analyses. To compute this phylogeny, a short-read assembly for the toxigenic *V. cholerae* O139 described in Chapter 4 was used, rather than the corresponding long-read assembly. Consequently, the duplication of VSP-1 was not detected in this isolate, although its duplication had been demonstrated repeatedly in Chapter 4. However, VSP-1 was found to be duplicated in the closed genome of MJ-1236, consistent with a previous report of this isolate [98].

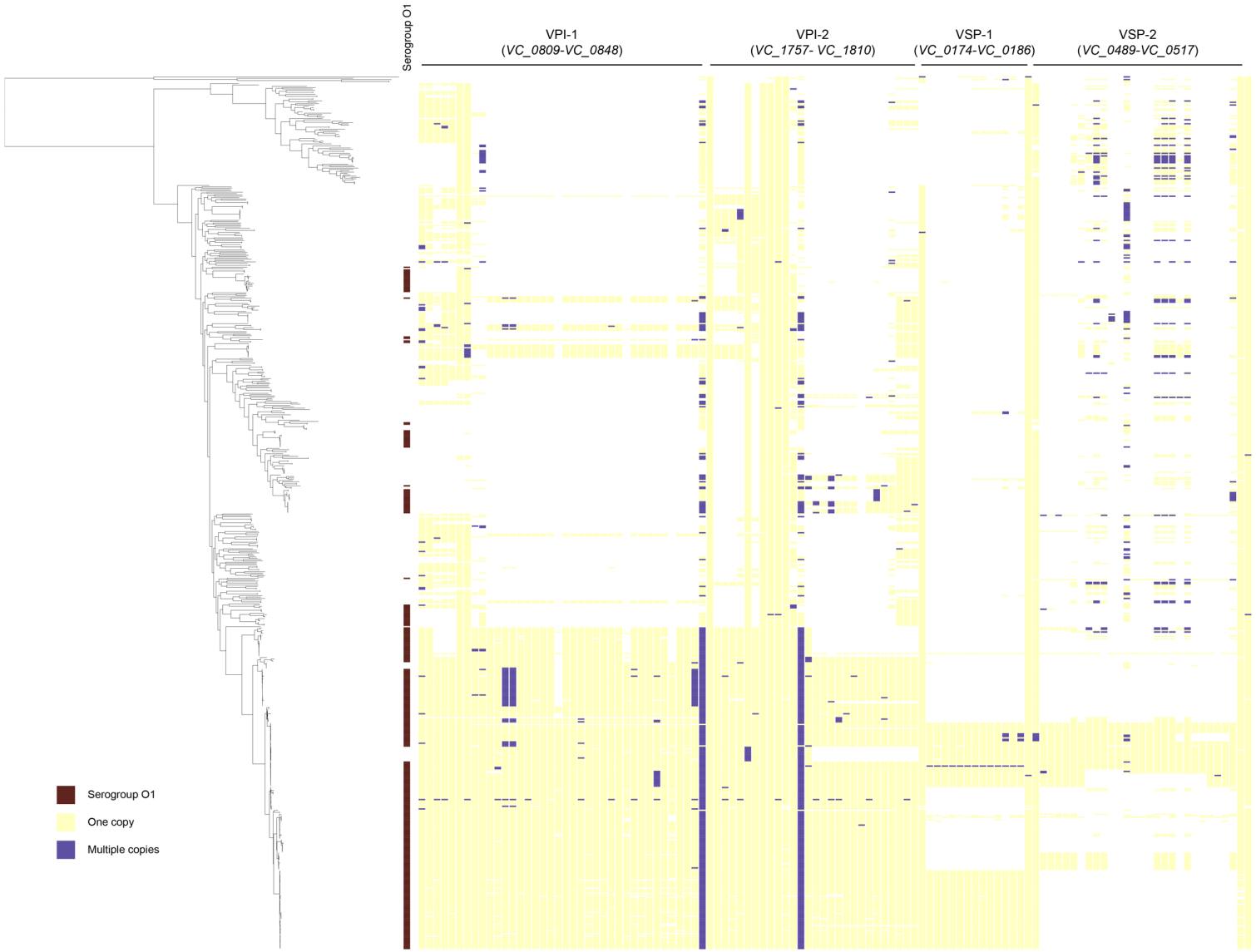


Figure 5.7 – Distribution of genes encoded by canonical pathogenicity islands in the *V. cholerae* pangenome. Legend on following page.

Figure 5.7 legend: The locus IDs from the N16961 reference genome for each pathogenicity island are indicated. Genes flanking these pathogenicity islands were also included, and are evident as genes at the beginning or end of each island which are present in nearly every isolate in the dataset. Serogroup assignment was taken from Figure 5.6. Since this plot reports the presence and absence of elements found in N16961, $tcpA^{El\ Tor}$ appears to be absent from genomes which harbour the $tcpA^{Classical}$ allele, though the remainder of VPI-1 genes are present in such genomes

5.3.6 – Plasmid and antimicrobial resistance gene distribution amongst V. cholerae

Continuing the exploration of these diverse genomes, all genome assemblies were scanned for the presence of antimicrobial resistance determinants and plasmid replicon sequences (see Methods). The results of this analysis are summarised in Figure 5.8. It was evident that isolates within the diverse clade (Figure 5.8, grey disc) harbour sequences homologous to the beta-lactamases bla_{CARB-7} and bla_{CARB-9}. These two genes have been discussed in Chapter 4; briefly, both are chromosomally-encoded within the integron on chromosome 2 [151, 152]. This is consistent with the lack of plasmid replicon sequences amongst these genomes (Figure 5.1, 5.8). In this analysis, NCTC 30 was determined to harbour bla_{CARB-7}, which corresponds to the functionally-validated $bla_{CARB-like}$ gene (section 4.3.9). Other observations that were consistent with previous reports include the presence of a *catB9* gene within 7PET [158], and the detection of *qnrVC* quinolone resistance genes amongst Chinese non-7PET genomes [396] (Figure 5.8).

It is important to note that the *tet(34)* sequence, predicted to be present in 650 of 651 isolates in this dataset, is homologous to a xanthine-guanine phosphoribosyltransferase (XPRT) gene present on the *V. cholerae* chromosome [59, 441]. This sequence was not detected in analyses for previous chapters, which made use of ARIBA to scan short-read data for resistance determinants (Chapter 3), and the Resfinder web interface, which scans assemblies using BLASTn (Chapter 4). Moreover, tetracycline resistance was not reported in the vast majority of phenotyped *V. cholerae* included in this dataset (Chapters 3, 4). Since this was the first time that ABRicate had been used to detect resistance genes in assemblies, it is possible that detection of *tet(34)* is an artefact of the chosen analysis method. Additional validation would be required before further inferences can be drawn from this observation.

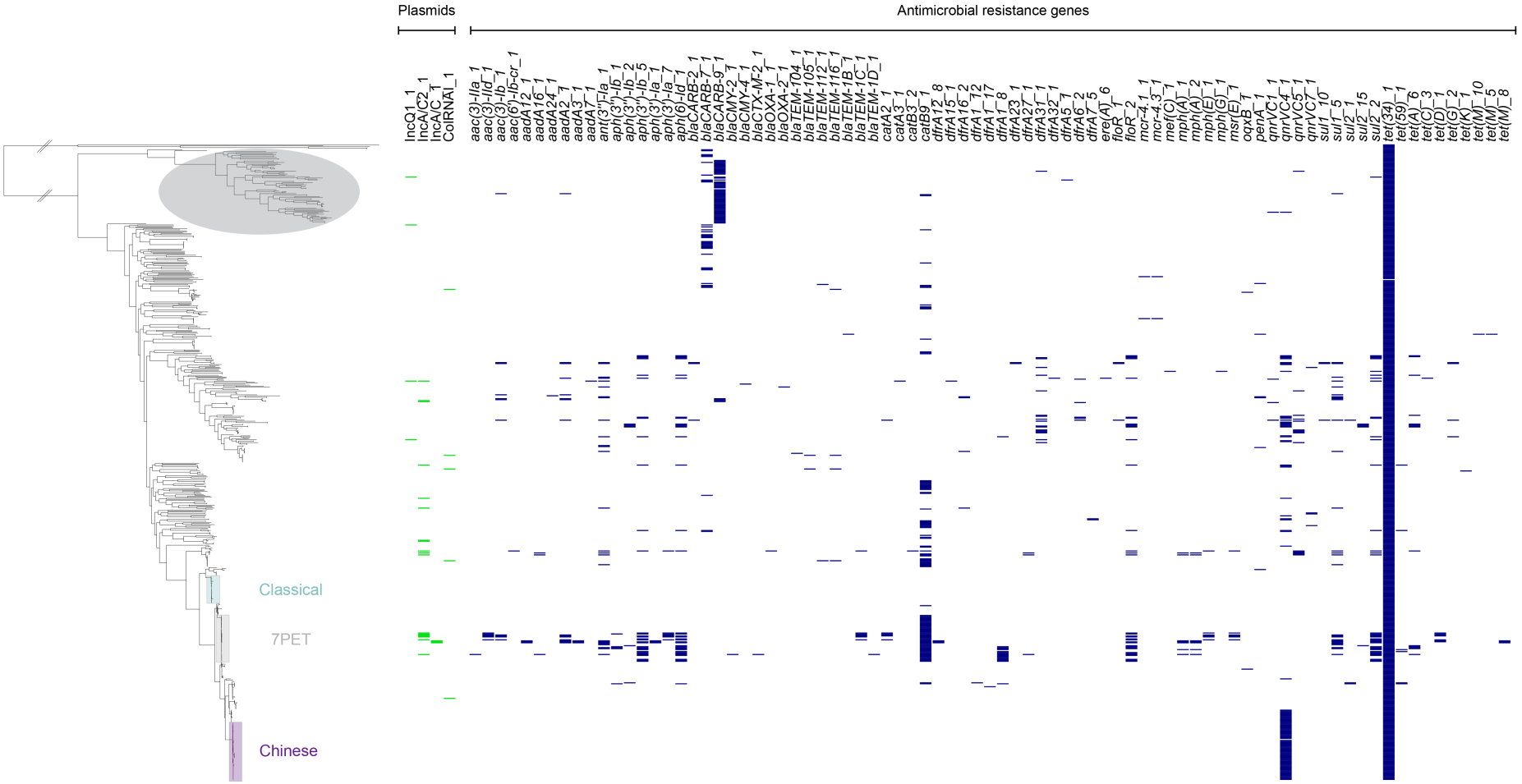


Figure 5.8 – Distribution of AMR genes and plasmid replicons within the *V. cholerae* phylogeny. Hatch marks denote branches which have been manually shortened for illustrative purposes. Assemblies were scanned for sequences of interest using ABRicate and the ResFinder/PlasmidFinder databases (see Methods). Plasmids, AMR genes >75% coverage and identity

From these data, it was evident that the majority of sequenced *V. cholerae* possessed very few antimicrobial resistance genes; 550 of the 651 genomes harboured two or fewer AMR determinants (Figure 5.9). The presence of a plasmid replicon is statistically significantly associated with a *V. cholerae* isolate harbouring three or more AMR genes (χ^2 (d.f. = 1, n = 650) = 75.84; $p < 0.00001$; Table 5.1). This suggests that the majority of AMR genes amongst this dataset are present on plasmids, of which there are very few types amongst these genetically heterogenous isolates.

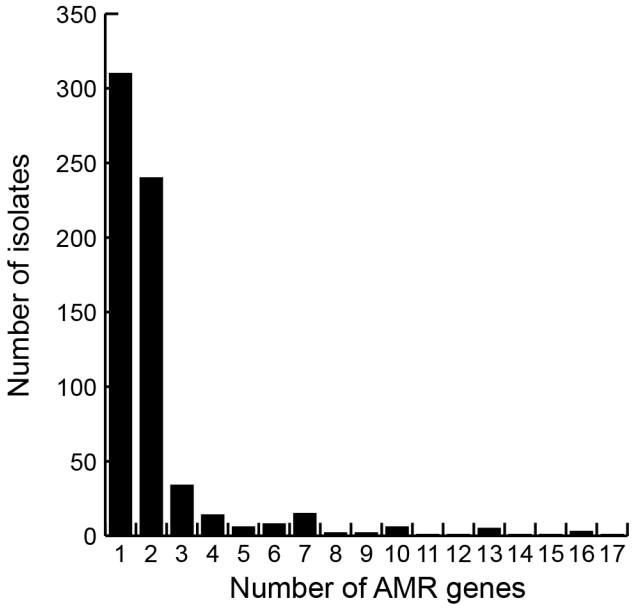


Figure 5.9 – The vast majority of *V. cholerae* isolates harbour two or fewer AMR genes. The potentially-dubious detection of *tet(34)* gene means that those isolates predicted to harbour one resistance gene are likely to be pan-susceptible.

	Plasmid Replicon		
AMR genes	No	Yes	Σ(rows)
≤ 2	542	8	550
≥ 3	79	21	100
Σ(columns)	621	29	650

Table 5.1 – χ^2 contingency table. Testing for association between the presence of a plasmid replicon (IncQ1, IncA/C1, IncA/C2, ColRNAI) and increased numbers of AMR genes.

It was evident that within this dataset were a number of serogroup O1 *V. cholerae*, many of which were part of a collection of *V. cholerae* supplied by NCTC. Many of these were the lineages local to Latin America that were previously-described [189], or were from a recent study of Chinese non-7PET *V. cholerae* O1 [396]. However, a number of non-7PET *V. cholerae* O1 were evident which were part of the NCTC collection of historical *V. cholerae* isolates. This prompted a closer investigation of the history and genomes of the NCTC *V. cholerae*, discussed below.

5.3.7 – Phylogenetic positions of historically-important NCTC isolates

As part of this thesis research, I spent time at PHE Colindale examining historical records for *V. cholerae* contained within the NCTC collection, and collating these data with the assistance of Jake Turnbull. Names and basic metadata for the 35 isolates included in this chapter are listed in Table 5.2. Following the detailed characterisation of the PacBio-assembly of NCTC 30 (Chapter 4), the phylogenetic position of all other *V. cholerae* that were in-stock and available from the NCTC collection was determined (Figure 5.1). All of these isolates were available as live cultures for experimentation under CL3 conditions, and were sequenced using both PacBio and Illumina technologies with gDNA prepared using the methodology optimised using NCTC 30 (Methods, section 2.2.5). All live isolates are denoted on Figure 5.1.

NCTC No.	Other strain references and names	Year of isolation	Internal isolate ID	Serogroup/Serotype from records	Phylogenetic lineage (Figure 5.1)	*In silico* serogroup
30	ATCC14735; MARTIN 1	1916	MJD382	Non-O1/O139 (O2)	-	Non-O1
3661	CN 5870; DOORENBOS 80	1931	MJD383	O1 El Tor	near-ELA5	O1
4693	JAPANESE ORIGINAL	pre-1936	MJD384	O1 Inaba	Classical	O1
4711	ATCC 14730; NANKING 32/123	Pre-1936	MJD385	O2	-	Non-O1/O139
4714	EL TOR 34-D 19	1934	MJD386	Non-O1/O139	-	Non-O1/O139
4715	ATCC 14731; CN 3426; ELTOR 34-D 23	1934	MJD387	O3	-	Non-O1/O139
4716	ATCC 14732; KASAULI 73	1932	MJD388	O4	-	Non-O1/O139
5395	ATCC 14734; IRAQ	1938	MJD389	O1 El Tor Ogawa	pre-7PET	O1
5596	SHANGHAI 10	pre-1939	MJD390	O1 Ogawa	Classical	O1
6585	SUBAMMA	pre-1944	MJD391	No data	Classical	O1
7258	EGYPT 109	pre-1948	MJD392	No data	Classical	O1
7260	EGYPT 117	pre-1948	MJD393	No data	Classical	O1
7270	HIKOJIMA	pre-1948	MJD394	O1 Inaba	Classical	O1
8021	ATCC 14035; RH 1094	pre-1950	MJD395	O1 Classical Ogawa	Classical	O1
8039	CAIRO 1A	pre-1950	MJD396	O1 Classical Inaba	Classical	O1
8040	726/575 A	pre-1950	MJD397	O1 Classical Inaba	Classical	O1
8041	757 AUTOAGGLUTINABLE	pre-1950	MJD398	O Rough	Classical	O1
8042	ATCC 14733; WDCM 00203	pre-1950	MJD399	Non-O1/O139	-	Non-O1/O139
8050	CALCUTTA	pre-1950	MJD400	No data	-	O1
8367	4 Z	pre-1952	MJD401	O1 Classical Ogawa	Classical	O1
8457	ATCC14033; CN5774; DO1930; RH1092	1930	MJD402	O1 El Tor	near-ELA5	O1
9420	CN 5789; TOR A	pre-1955	MJD403	El Tor	pre-7PET	O1
9421	CN 5790; TOR 1	pre-1955	MJD404	O1 El Tor Ogawa	pre-7PET	O1
9422	CN 5871; TOR 8	pre-1955	MJD405	O1 Inaba	near-ELA5	O1
9423	TOR 31	pre-1955	MJD406	No data	pre-7PET	O1
10255	CN 5745; 20109	1961	MJD365	O1 El Tor Ogawa	7PET	O1
10256	CN 5748; 20111	1961	MJD366	O1 El Tor Ogawa	7PET	O1
10732	CN 3534; 384/52	1952	MJD367	O1 Classical Inaba	Classical	O1
10733	AJ 1592; CN 3539; I-5; 586/52	1952	MJD368	O1 Classical Ogawa	Classical	O1
10954	1330	1973	MJD369	El Tor	7PET	O1
11090	NCIB 9341; 1077	pre-1950	MJD370	No data	-	O1
11348	VL3029; WDCM00136, CCUG67718, DSM101014	pre-1981	MJD372	O24	-	Non-O1/O139
11500	VL 7050	pre-1983	MJD373	No data	-	Non-O1/O139
11643	VL 4944	pre-1985	MJD380	No data	-	Non-O1/O139
12946	ATCC 51395; MO3	1993	MJD381	O139	7PET (O139)	O139

Table 5.2 – NCTC *V. cholerae* sequenced for this thesis research. Excluded from this list is one sequenced isolate which was found to be a member of the *Aeromonas* genus upon analysis of the genome assembly.

5.3.7.1 – Pandemic NCTC isolates

Of the 35 NCTC isolates that were included in this analysis, four were found to be members of 7PET (Figure 5.10). One of these was NCTC 12946, also known as MO3, a serogroup O139 strain isolated at the beginning of the *V. cholerae* O139 epidemic in 1993 [436]. Genomic analysis confirmed that this isolate possessed serogroup O139 LPS genes, was toxigenic, and was phylogenetically positioned amongst previously-sequenced toxigenic *V. cholerae* O139 within 7PET (Figure 5.10). NCTC 10255 and 10256 were isolated from patients during an outbreak of "paracholera" in Hong Kong during 1961 [358]. Vella demonstrated that these El Tor Vibrios were highly virulent in a mouse model, and that these isolates showed some ability to immunise mice against re-infection [358]. NCTC 10954 has been used historically as a positive control strain for haemagglutination in *V. cholerae* biotyping [40].

Thirteen isolates were members of the Classical lineage (Figure 5.10). Vella compared NCTC 10255 and 10256 to a "true cholera" vibrio, NCTC 7260, the most virulent isolate available at the time [358] (Table 5.1; Figure 5.10). The finding that this is a Classical isolate is logical, given that Vella had referred to this as a "true cholera" *Vibrio* when compared to El Tor isolates [358]. Other isolates of interest that were members of the Classical lineage include NCTC 8367 (strain 4Z), which has been used as a source of neuraminidase ("receptor destroying enzyme") [135, 442]. Neuraminidase is encoded by *nanH*, part of VPI-2 [134] (section 1.2.5), a pathogenicity island found in Classical *V. cholerae* [54, 133, 134] and in this isolate (Figure 5.7). Characterising the phylogenetic position of NCTC 8021 is particularly important because this isolate is the *Vibrio cholerae* type strain and the neotype of the *V. cholerae* species [443]. In 1965, Hugh compared the biochemical phenotypes of 258 *V. cholerae* isolates to that of NCTC 8021 [444], and requested an opinion from the Judicial Commission of the International Committee on Bacteriological Nomenclature that this strain be considered the *V. cholerae* neotype [209]. *V. cholerae* appeared on the Approved List of Bacterial Names in 1980 [35], and Hugh's work was cited as the description of the *V. cholerae* species [35, 444]. NCTC 8021 has been used as a reference standard for the negative haemagglutination phenotype used to biotype Classical *V. cholerae* [40]. It is reassuring that this type strain is a member of the Classical lineage, given that NCTC 8021 was designated a type strain during the 1960s, at a time when Classical *V. cholerae* was still considered to be the sole aetiological agent of cholera [1, 36]. Thus, the type strain chosen was indeed biochemically, microbiologically, and genomically, a representative of the aetiological agent of Asiatic cholera.

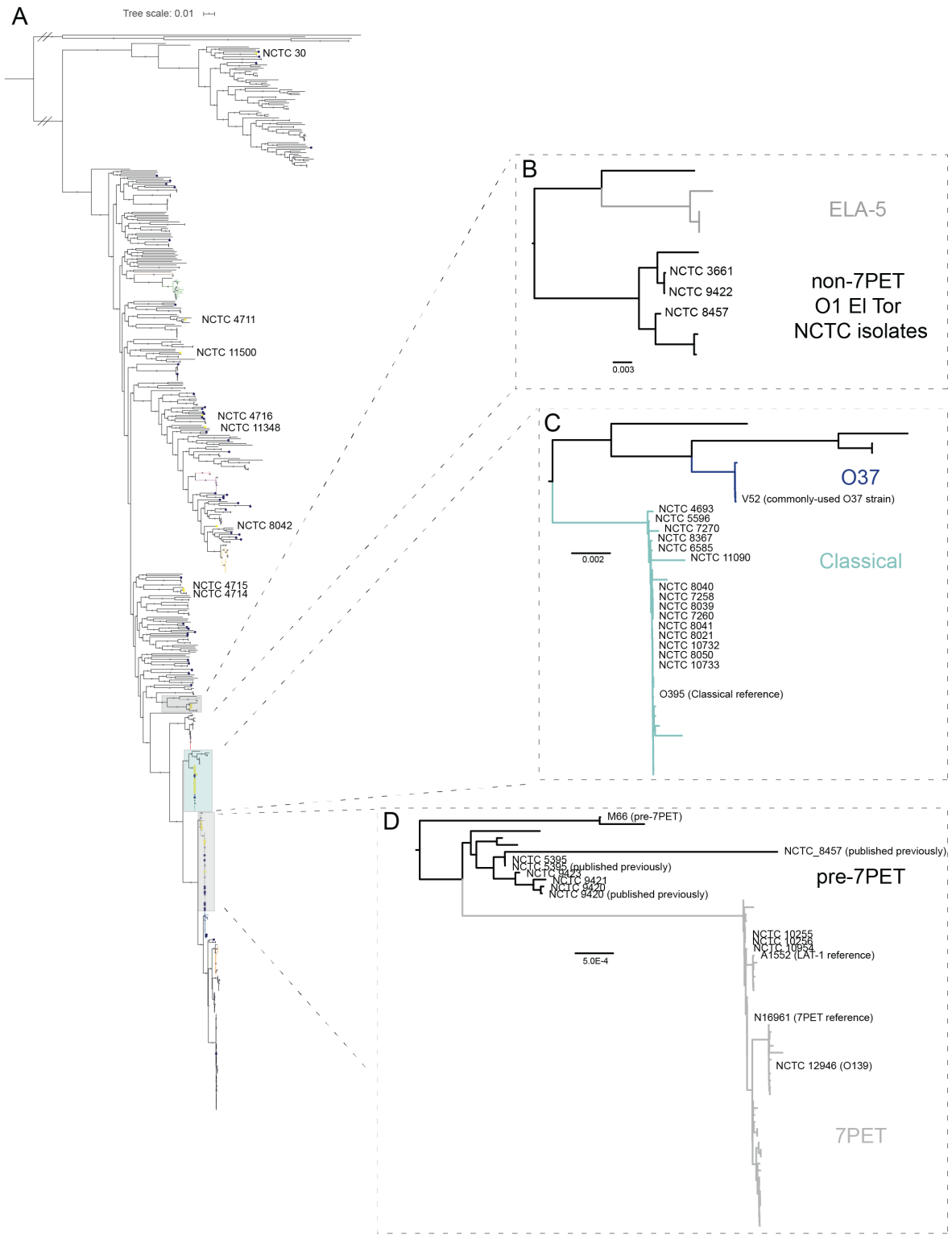


Figure 5.10 – A *V. cholerae* phylogenetic tree annotated with the names and IDs of NCTC isolates. The phylogeny presented in Figure 5.1 is re-drawn and the position of NCTC isolates indicated. Where necessary for illustrative clarity, sub-trees were extracted from the original phylogeny and visualised separately (Figtree v1.4.3). The names of NCTC isolates and key reference sequences have been retained. Hatch marks indicate branches that were manually shortened for illustrative purposes.

5.3.7.2 – Non-pandemic NCTC V. cholerae *O1*

Of interest was a set of three *V. cholerae* O1 which were not part of either the 7PET or Classical lineages. NCTC 3661, 8457, and 9422 were all recorded as being serogroup O1 in NCTC records (Table 5.2). NCTC 3661 and 8457 were also recorded as being biotype El Tor (Table 5.2). Genomic analysis confirmed that these three isolates harboured the genes necessary to make these serogroup O1, but lacked VPI and VSP pathogenicity islands and CTXφ, though they were predicted to harbour T3SS-2α (Figure 5.6, 5.7, 5.8). In the absence of any further clinical metadata, it is impossible to know whether the T3SS-2α predicted to be present in these isolates is either functional or contributed to causing disease in a patient.

These three isolates were particularly interesting because of their history. NCTC 3661, Doorenbos 80, was isolated in 1931 from a healthy Mecca pilgrim [41]. Gardner and Ventakraman described this as having 'typical' biochemical characteristics, being of O-subrgoup I, and to be capable of haemolysing goat erythrocytes [41]. In subsequent studies, NCTC 3661 was also shown to be haemolytic on sheep blood agar, produced a positive Voges-Proskauer test result, and was positive in a Grieg test for haemolysis [445]. However, this strain has been shown to be sensitive to group IV cholera bacteriophage, a phage to which El Tor isolates tend to be resistant [446].

NCTC received a batch deposition of four *V. cholerae* from A.H. Wahba, which were received simultaneously from Cairo in 1953 (NCTC records). At the time, these isolates were named Tor A, Tor 1, Tor 8 and Tor 31 (Table 5.2), and were simply described as "El Tor strains" (NCTC records). These were accessioned into the NCTC collection as NCTC 9420-9423. A recent study reported genome sequences for NCTC 9420 and NCTC 5395, and found that both of these sequences were basal to the 7PET lineage [213], dubbed "pre-7PET" here. These sequences are included in the phylogeny presented in Figure 5.10. It has been stated that the properties of NCTC 9420, as well as NCTC 8457 and 5395, are likely to resemble those of the original El Tor biotype *V. cholerae* described by Gotschlich, which tended to be isolated from asymptomatic patients [211, 213]. For this PhD project, NCTC 9420, 9421 and 9422 were sequenced. Although NCTC 9420 and 9421 are basal to 7PET, isolate NCTC 9422 sits adjacent to NCTC 3661 (Figure 5.10). This strongly suggests that the El Tor phenotypes, which led NCTC 9420-9423 to be isolated and deposited with NCTC, are consistent between the pre-7PET isolates and the phylogenetically-unrelated NCTC 9422.

A closer inspection of the NCTC 3661 and NCTC 9422 genome sequences demonstrated that both isolates were predicted to harbour IncA/C2 plasmids (Figures 5.1, 5.8). This was surprising, because when the assemblies for both of these isolates were scanned for antimicrobial resistance determinants, no putative resistance genes were detected (Figure 5.8). To address this, a combination of PacBio RSII long-reads and Illumina short-reads were used to produce hybrid assemblies for these genomes. Unicycler, with 'conservative' settings, was used to minimise the likelihood that contigs would be merged spuriously (see Methods). A ~130 kb circularised contig was identified in the hybrid assemblies of both genomes. Visualisation of the De Brujin graph from the corrected, circularised assembly confirmed that this element was independent of the two *V. cholerae* chromosomes (see Figure 5.11 for representative result from NCTC 3661), and this contig contained the IncA/C2 replicon. No antimicrobial resistance genes were detected on this contig.

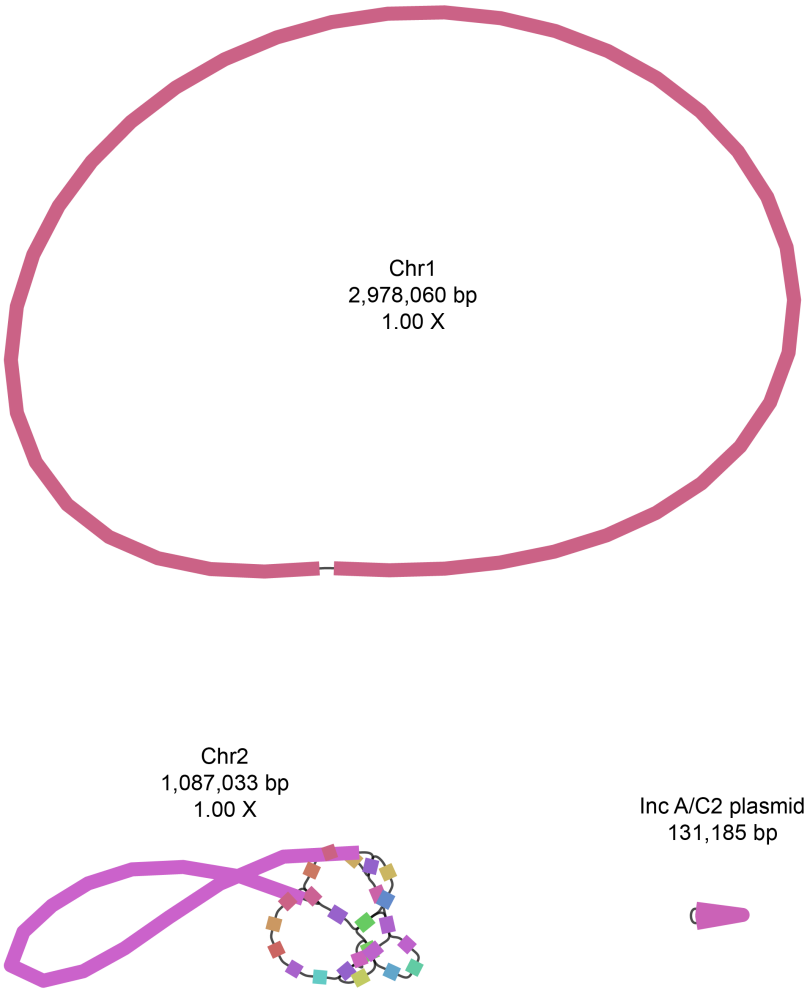


Figure 5.11 – Visualisation of the De Brujin graph for the polished, rotated hybrid assembly for NCTC 3661. The integron on chromosome 2 was not completely assembled; this is a consequence of running Unicycler on 'conservative' settings. An identical plasmid of 131,185 bp was assembled from NCTC 9422.

Comparative genomics demonstrated the similarity between this putative IncA/C2 plasmid (pNCTC3661) and previously-published *V. cholerae* plasmids from the same family, including the rare MDR IncA/C2 plasmid detected in Argentinian LAT-1 *V. cholerae* described in Chapter 3 (Figure 5.12). It also highlighted that the variable regions of the plasmid backbone, which can harbour antimicrobial resistance genes, are much shorter in this drug-sensitive plasmid compared to plasmids encoding resistance determinants (Figure 5.12).

5.3.7.1 – NCTC 8457

The first report of NCTC 8457 was made by Doorenbos and Kop, in 1951, in which paper the isolate is recorded as having been isolated 20 years prior to the publication [447]. This strain is serogroup O1, biotype El Tor, and non-toxigenic [213]. On this basis, Hugh argued that NCTC 8457 should be the neotype of *Vibrio* eltor Pribram (*i.e.*, El Tor *V. cholerae*), since this isolate displayed all of the characteristics of an El Tor vibrio [210]. A genome sequence for NCTC 8457 was reported in 2009 [54] but was first sequenced in 2006 at TIGR (accession # NZ_AAWD00000000.1). This sequence has subsequently been used as an exemplar of the "pre-seventh pandemic" group of *V. cholerae* [54, 213]. In particular, this strain was used alongside other historical isolates to make inferences about the evolution of the lineage now dubbed 7PET [213].

Unlike our sequence data for NCTC 9420 and 5395, the phylogenetic position of our sequenced isolate of NCTC 8457 was not consistent with that of the previously-reported genome (Figure 5.10). Rather than co-clustering with the previously-sequenced NCTC 8457 isolate as did our re-sequenced NCTC 5395 and 9420 (Figure 5.10D), our stock of NCTC 8457 clustered with the two serogroup O1 biotype El Tor isolates NCTC 9422 and 3661 (Figure 5.10B). This was a surprising result, but was replicated when gDNA from an independent batch of NCTC 8457 (DNA supplied by M.A. Fazal at NCTC) was sequenced and similarly analysed. This suggested that there may be discrepancies between the previously-sequenced culture of NCTC 8457 and the stocks which were sequenced for this PhD project.

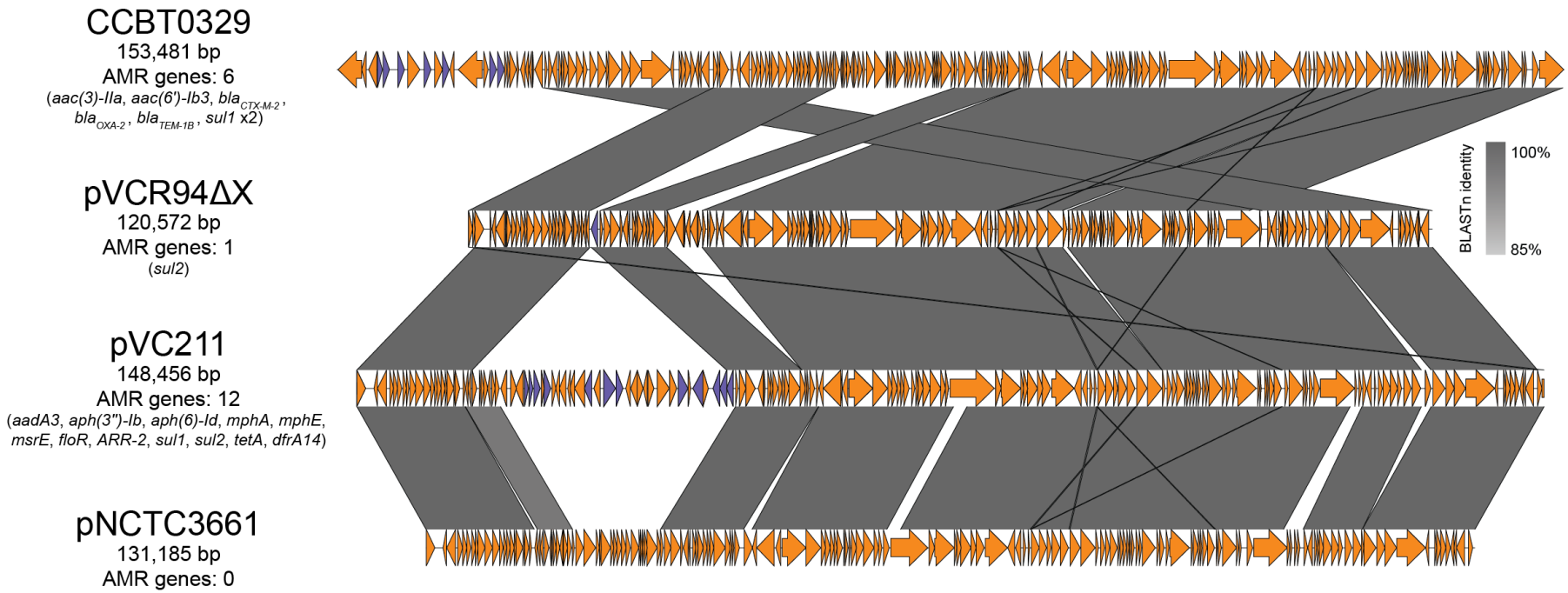


Figure 5.12 – Comparing pNCTC3661 to three *V. cholerae* IncA/C2 plasmids. Included in this comparison is the Argentinian IncA/C2 plasmid described in Chapter 3 (Figure 3.20).

In light of this apparent discrepancy, a number of confirmatory analyses were performed to validate our genomic observations, and to provide support for the phylogenetic position of this sequence. Firstly, it was confirmed *in silico* that NCTC 8457 was likely to be of serogroup O1 (Figure 5.6). This is fully-consistent with previous reports [54]. Similarly, a manual inspection of the genome confirmed that CTXφ and the genes encoding cholera toxin could not be detected in this genome (Figure 5.7); again, consistent with previous characterisations of the isolate.

Interestingly, a conservatively-generated hybrid assembly for NCTC 8457 suggested the presence of a small circular element in this assembly, which was independent of the two chromosomes (Figure 5.13). No plasmids were detected in the hybrid assembly using the Plasmidfinder website or in the earlier analysis of the short-read assembly (Figure 5.8).

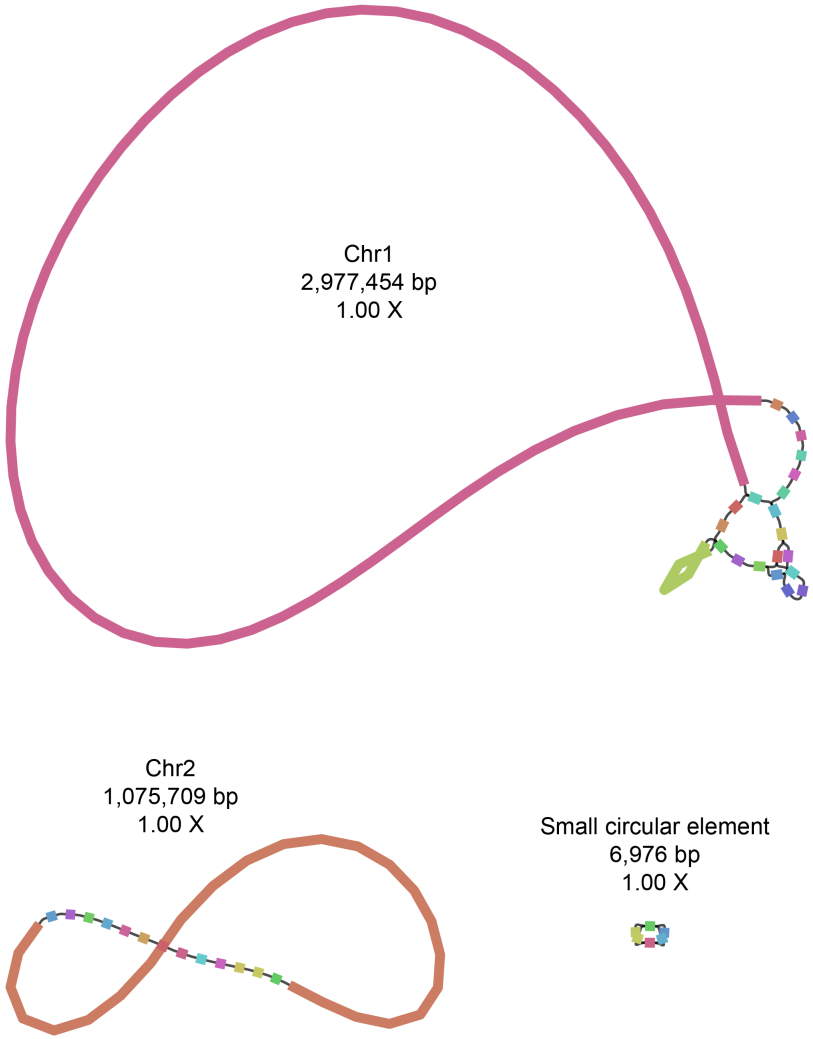


Figure 5.13 – Visualisation of the De Brujin graph for the NCTC 8457 hybrid assembly. Figure produced using Bandage and the polished assembly graph generated by Unicycler.

The six contigs that comprised this 6.9 kb element were extracted from the hybrid assembly, and their concatenated sequence was used to query the nr database using BLASTn. The most similar sequence to this element was a 3.8 kb plasmid called pVC (accession # AY423429), the sequence of which covered the query from NCTC 8457 by 69%, with 98.99% nucleotide identity. pVC is a cryptic plasmid which was isolated from an environmentally-derived non-O1/O139 *V. cholerae* isolate dubbed MP-1 [448]. This manuscript reported that four CDSs of unknown function were identifiable on pVC, that attempts to cure pVC from MP-1 were unsuccessful, and that pVC was likely to replicate *via* theta replication [448, 449]. The replication origin of pVC was identified in this work [448], and this sequence is shared between pVC and the putative plasmid from NCTC 8457.

In order to confirm further whether the 6.9 kb element in NCTC 8457 might be an extrachromosomal plasmid, plasmid extractions were performed using stationary-phase cultures of this strain and others (alkaline lysis using Qiagen Midiprep kits; Methods, section 2.2.14). A small DNA element was successfully extracted from cultures of NCTC 8457 (stock ID MJD402; Figure 5.14), further suggesting that this isolate harbours a small plasmid that replicates extrachromosomally. Additional support for these observations was obtained while reviewing the literature surrounding NCTC 8457 - a previous publication had reported that NCTC 8457 harboured a small cryptic plasmid [450]. In future work, excising this DNA element and using a combination of Sanger sequencing and primer walking may allow for the identity and sequence of this element to be confirmed.

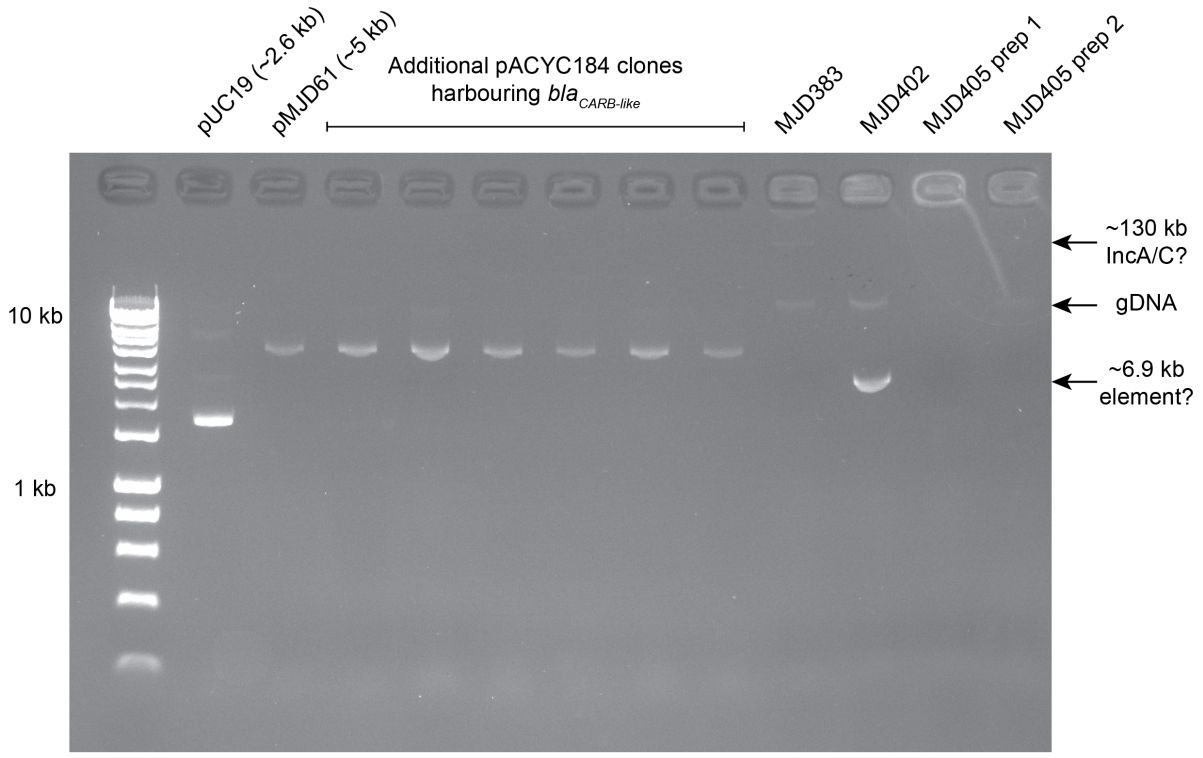


Figure 5.14 – Gel of plasmid preps from *V. cholerae* and *E. coli*. The plasmid preparations presented here were made from different organisms (*E. coli* and *V. cholerae*). Plasmids of known size from *E. coli* (lanes 1-8) were extracted from 2 ml stationary-phase overnight cultures using a Qiagen miniprep kit (Methods). Plasmids were extracted from 25 ml cultures of *V. cholerae* (lanes 9-12) using a Qiagen midiprep kit (Methods). Plasmids were not linearised before electrophoresis and host organisms were not consistent; hence, sizes are approximate. Ladder: Hyperladder 1 kb.

5.3.8 – Biotype determinants

The collection of historically-important isolates described above led us to investigate the basis of biotyping across this phylogeny. We chose to do this because (a) a number of Classical genomes had been added to the collection, and (b) three O1 El Tor isolates had been characterised which were distantly related to 7PET. Understanding the distribution of biotype-determining mutations across the species is an important step towards understanding what the "default" biotype would be for a newly-emerging O1 lineage, though it is accepted that biotyping reactions are uninformative when used on non-O1 *V. cholerae* [206]. The three principal phenotypes used to biotype *V. cholerae* O1 are sensitivity to polymyxin B, the Voges-Proskauer test, and haemolysis [40, 206]. The molecular and genetic bases of each of these phenotypes have been characterised, and were investigated across the phylogeny.

5.3.8.1 – Voges-Proskauer test and acetoin biosynthesis in V. cholerae

The Voges-Proskauer test detects acetoin in bacterial cultures [451]. Acetoin is produced as an intermediate molecule by *V. cholerae* during the fermentation of glucose to 2,3-butanediol [452]. El Tor *V. cholerae* produce acetoin and test positive when subjected to the Voges-Proskauer test, and classical *V. cholerae* do not produce acetoin, producing negative Voges-Proskauer test results [206]. The inability of classical *V. cholerae* to ferment glucose (*via* acetoin) means that classical biotype strains exhibit a severe growth defect when cultured on minimal media containing glucose as a sole carbon source, acidifying their growth media [452].

It has been shown that the AphA transcription factor negatively regulates acetoin production in *V. cholerae* [453]. This occurs because AphA directly repressed the *alsD* gene and the gene encoding the AlsR transcriptional regulator [453]. Additionally, the ability to produce (p)ppGpp has been shown to regulate acetoin production; this is independent of AphA [454]. It has also been shown that HapR, the master regulator of *V. cholerae* quorum-sensing responses, represses *aphA* when cultures reach a high cell density [455]. Thus, at high cell density, HapR represses *aphA*, thereby de-repressing the genes whose products are required to ferment acetoin (Figure 5.15) [453]. It has been shown that a single G→T mutation at position -77 in the *aphA* promoter ablates the ability of HapR to bind to, and repress, transcription of *aphA* [455]. It has been demonstrated that the -77T mutation in P_{aphA} is present in the Classical *V. cholerae* reference strain, O395, and that the -77G allele is present in 7PET isolate C6706 [455].

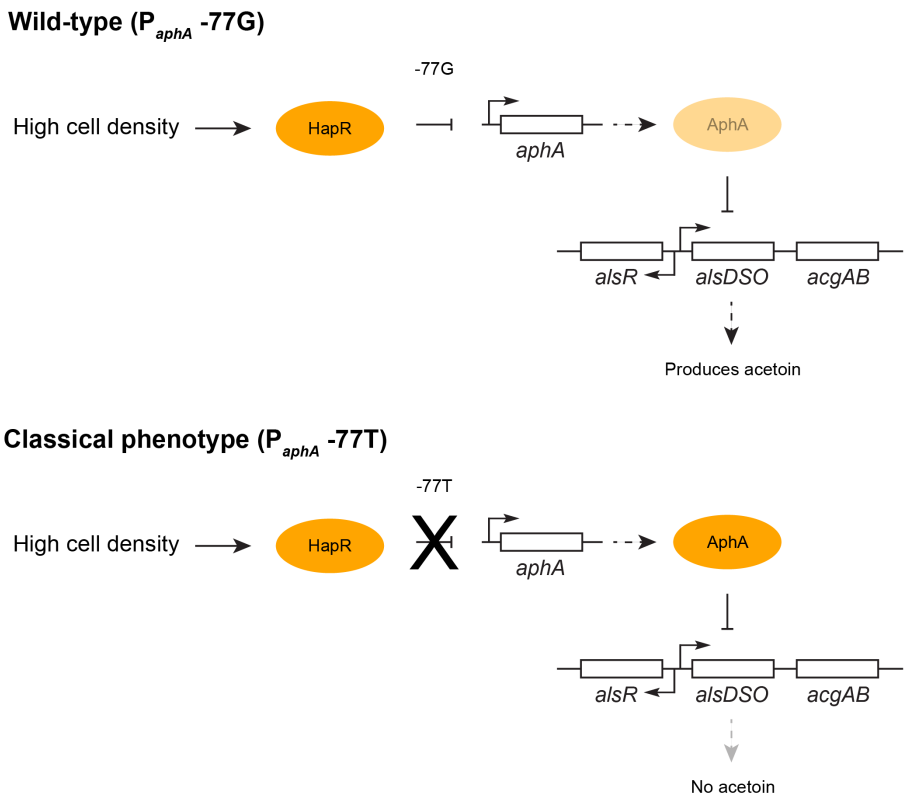


Figure 5.15 – Overview of the effects of P_{aphA} alleles on the expression of acetoin metabolism genes. At high cell density, HapR levels accumulate, and repress transcription of *aphA* if HapR binds to P_{aphA} (only if -77G allele is present). Diminished rates of *aphA* transcription and AphA translation leads to a concomitant de-repression of the *als* genes required to ferment acetoin.

In order to determine the P_{aphA} variants present in each isolate included in this dataset, *in silico* PCR was used to extract the *aphA* promoter sequence from each genome assembly (parameters listed in Methods, section 2.1.17). P_{aphA} was successfully extracted from 632 of the 651 sequences in the dataset. These sequences were aligned and compared to identify the variable sites in the promoter (Figure 5.16).

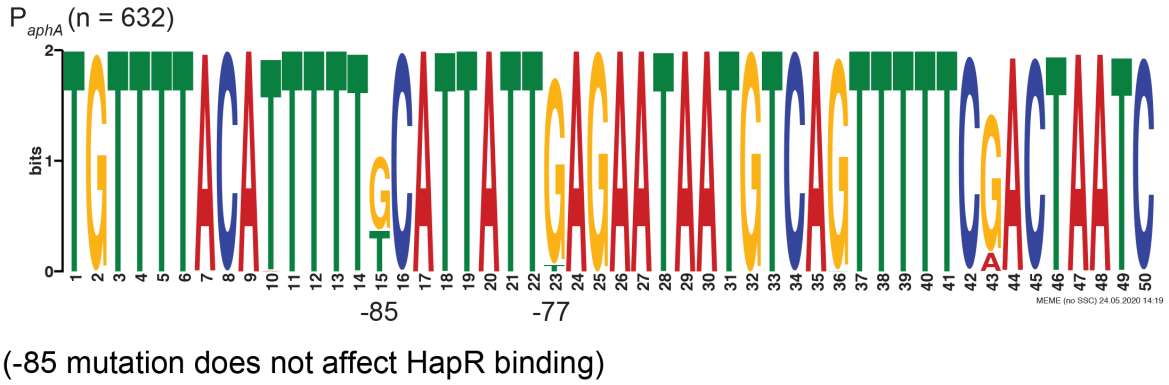


Figure 5.16 – P_{aphA} motif generated from a Clustal Omega alignment of 632 P_{aphA} sequences extracted using *in silico* PCR. The 50 bp region of interest was extracted from the "amplified" sequence using trimal. The MEME web server was used to produce the motif figure.

Two variable sites were visible in the P_{aphA} sequence that had been previously characterised – these were position -85, known not to influence HapR binding [455], and position -77, mutation of which from G to T abolishes HapR binding to P_{aphA} [455].

5.3.8.2 – Cholera toxin expression: an additional contrasting phenotype between classical and El Tor biotype isolates

The original study of *aphA* regulation by HapR was carried out in order to understand how HapR contributed to virulence gene expression in *V. cholerae* through AphA, and how this varied in a biotype-specific manner [455]. Canonically, classical biotype *V. cholerae* express the *toxR* virulence regulon *in vitro* at much higher levels than do El Tor isolates, which need to be cultured under stringent conditions to induce expression of virulence genes and CT *in vitro* [456, 457]. At low cell density, AphA and AphB co-ordinately activate the *tcpPH* promoter, and the binding of AphA to P_{tcpPH} enhances the binding of AphB to P_{tcpPH} [458, 459]. Resultant increases in TcpP and TcpH levels lead to an upregulation of *toxT* and the activation of virulence gene expression [458]. In classical isolates, the -77T mutation in P_{aphA} prevents *aphA* from being repressed by HapR; therefore, in classical isolates, the virulence regulon is not repressed at high cell density. In El Tor isolates, P_{aphA} is sensitive to HapR repression; therefore, at high cell density, AphA-mediated activation of P_{tcpPH} is reduced [458].

The differential activation of P_{tcpPH} by AphB is known to be necessary and sufficient to drive the classical and El Tor differences in virulence gene expression [460]. This has been shown

to be due to a single mutation in P_{tcpPH} ; the binding affinity of AphB for this promoter is tenfold higher if an A is present at position -65 [459]. In the O395 classical isolate, an A is present at -65, and in C6706, a G is present at the equivalent position (-66). Mutagenesis of both genetic backgrounds has shown that these alleles are sufficient to confer an El Tor or a classical virulence gene expression phenotype on *V. cholerae* [459]. This is summarised in Figure 5.17.

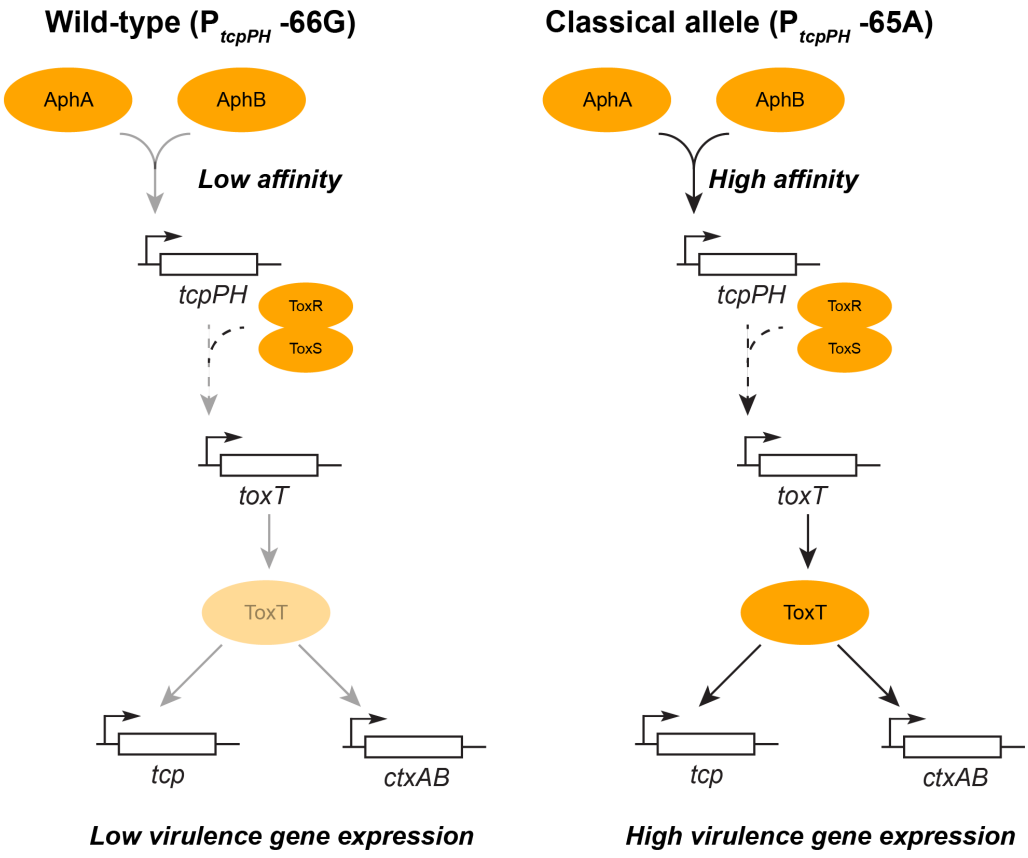


Figure 5.17 – Overview of the effects of P_{tcpPH} alleles on virulence gene expression. AphB (using AphA as a binding partner) has greatly enhanced affinity for P_{tcpPH} if an A is present at position -65/66. AphA/AphB binding to P_{tcpPH} activates expression of *tcpPH* and elevates intracellular levels of TcpP and TcpH. These, together with ToxRS, activate expression of *toxT*, thereby directly activating the expression of virulence genes. Conversely, in an "El Tor" background (P_{tcpPH} -66G), AphA and AphB have a lower affinity for P_{tcpPH}; therefore, virulence gene expression is lower.

It is important to consider mutations in both P_{tcpPH} and P_{aphA} at once, because together they couple the expression of virulence genes to quorum-sensing (Figure 5.18). In a Classical genetic background such as that of O395, *aphA* is rendered insensitive to repression by HapR. Therefore, regardless of cell density, *aphA* will be transcribed and AphA will be translated. Combined with the increased affinity for P_{tcpPH} experienced by AphA and AphB in this

background, this will lead to maximal expression of the *toxR*/*toxT* virulence regulon (Figure 5.18).

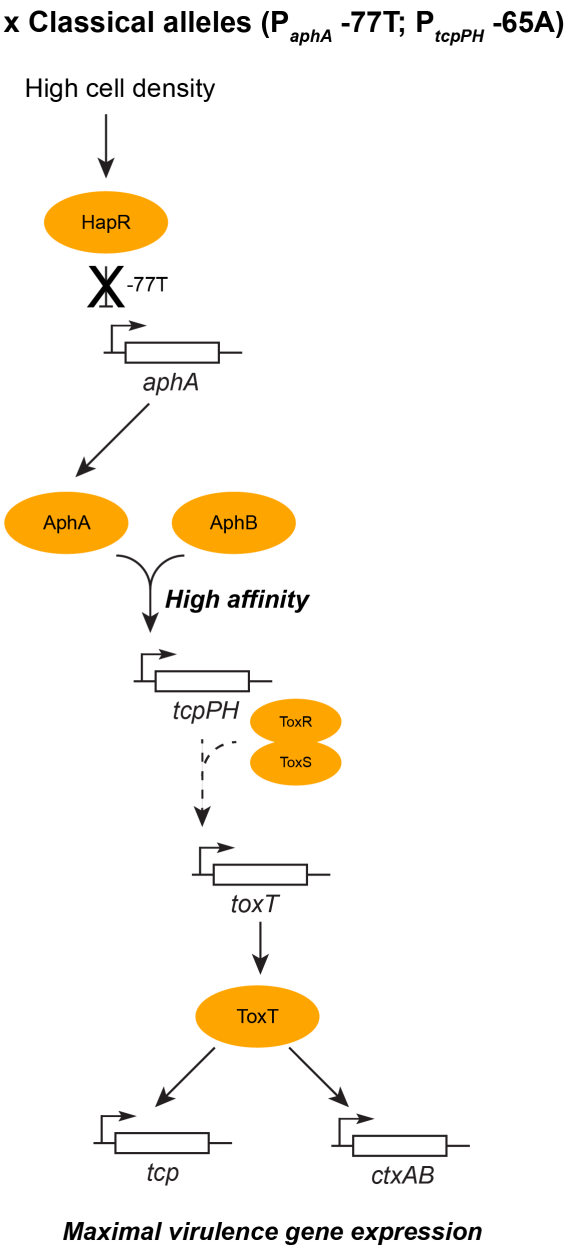


Figure 5.18 – Coordinate effects of Classical P_{tcpPH} and P_{aphA} alleles on virulence gene expression. The presence of both promoter alleles is predicted to lead to very high expression of the ToxT regulon.

In a similar manner to the study of P_{aphA}, the P_{tcpPH} sequence was extracted from all genomes in the dataset that harboured VPI-1 (258 isolates total), and compared to identify the variable sites in this promoter. This is illustrated in Figure 5.19.

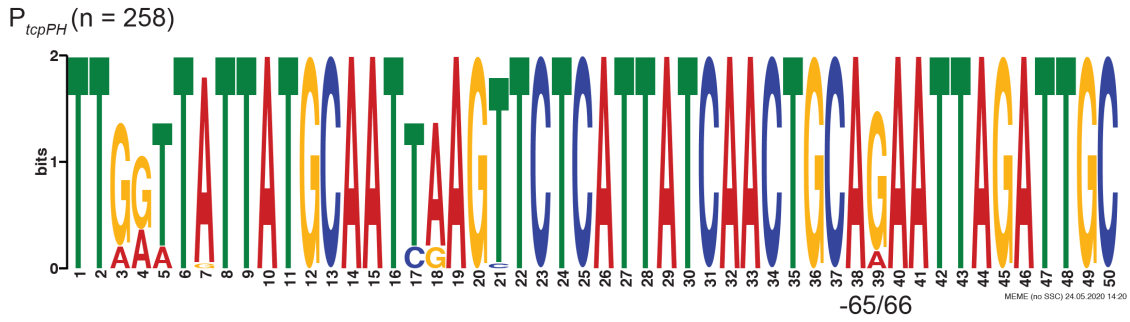


Figure 5.19 – P_{tcpPH} motif generated from a Clustal Omega alignment of 258 P_{tcpPH} sequences extracted using *in silico* PCR. The 50 bp region of interest was extracted from the "amplified" sequence using trimal. The MEME web server was used to produce the motif figure.

The genotype of P_{aphA} and P_{tcpPH} was determined for each genome in the dataset and mapped against the phylogeny, to determine the distribution of these variants (Figure 5.20). It was immediately apparent that both P_{aphA} -77T and P_{tcpPH} -65A were exclusively detected in Classical lineage *V. cholerae*, whereas the P_{aphA} -85 position was variable across the phylogeny.

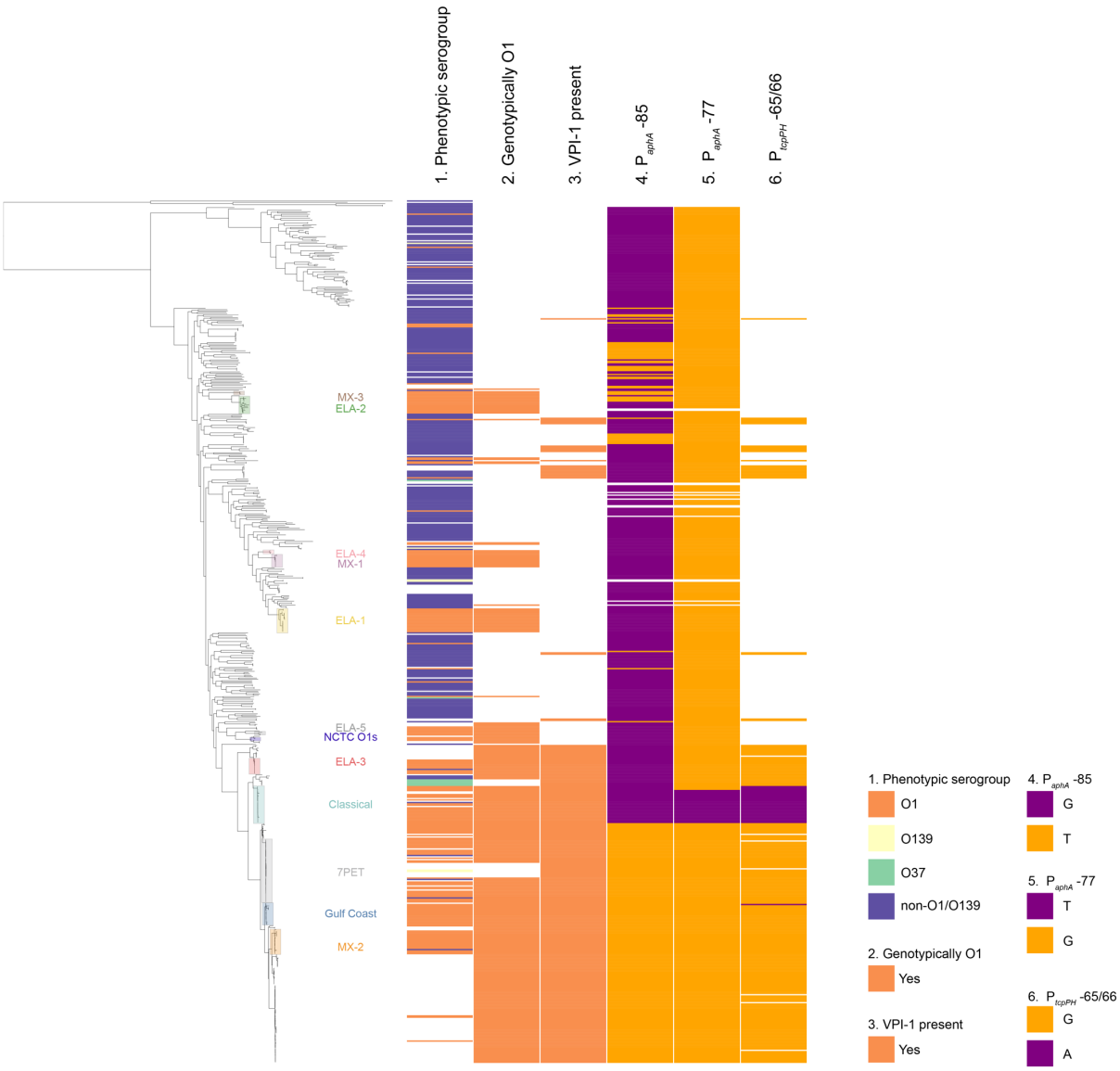


Figure 5.20 – Distribution of P_{aphA} and P_{tcpPH} allelic variants across the *V. cholerae* phylogeny. Isolates that were phenotypically serogroup O1 (for which those data were available) are indicated, as are those determined computationally to be serogroup O1. Lineages of *V. cholerae* O1 are indicated. Since *tcpPH* genes are harboured on VPI-1, P_{tcpPH} could only be detected in genomes containing VPI-1.

These data strongly indicate that genetic determinants which confer classical phenotypes (acetoin production and virulence gene expression) are almost exclusively in isolates that belong to the Classical lineage. Additional biotyping determinants were similarly studied, to determine whether this applies just to these promoters, or represents a more general observation.

5.3.8.3 – Polymyxin B sensitivity and haemolysis

El Tor biotype *V. cholerae* are classified as being resistant to polymyxin B – specifically, for an El Tor strain, a zone of inhibition will not be visible around a disc containing 50 units of polymyxin B [206]. In contrast, classical *V. cholerae* are sensitive to this concentration of polymyxin B and a zone of inhibition will be visible around the disc (the size of the zone of inhibition is not important in this specific assay) [206]. Briefly, the *almEFG* operon encodes three proteins which are responsible for catalysing the transfer of glycine residues to lipid A in El Tor *V. cholerae*, rendering them resistant to polymyxin B [461]. In a Classical isolate (O395), it was observed that the *almF* gene was disrupted, and that this mutation was responsible for the inability of this strain to resist polymyxin B – expression of an intact *almEFG* operon *in trans* dramatically elevated polymyxin B resistance in O395 [461].

Haemolysis in *V. cholerae* is mediated by the secreted haemolysin HlyA, which is processed into an active form by proteolysis [462, 463]. As mentioned in the Introduction (section 1.3.1.4), El Tor *V. cholerae* are characteristically haemolytic, and classical isolates are not [40, 206]. Classical isolates have been shown to be non-haemolytic as a consequence of a frameshift mutation in *hlyA* [464]. The hybrid strains of El Tor *V. cholerae* have been reported to be non-haemolytic, or haemolytic with a range of phenotypes [222]. Recent work suggests that multiple independent mutations in *hlyA* can be detected amongst hybrid *V. cholerae* [221]. It should also be noted that haemolysis is recognised to be an unreliable phenotype by which to characterise *V. cholerae* [206].

The state of *almEFG* and *hlyA* were determined across the phylogeny (Figure 5.21). Again, it was immediately evident that the mutations that confer a canonical classical biotyping phenotype – non-haemolysis due to an *hlyA* frameshift and sensitivity to polymyxin B due to an *almF* disruption – are both exclusive to the Classical lineage of *V. cholerae* (Figure 5.21).

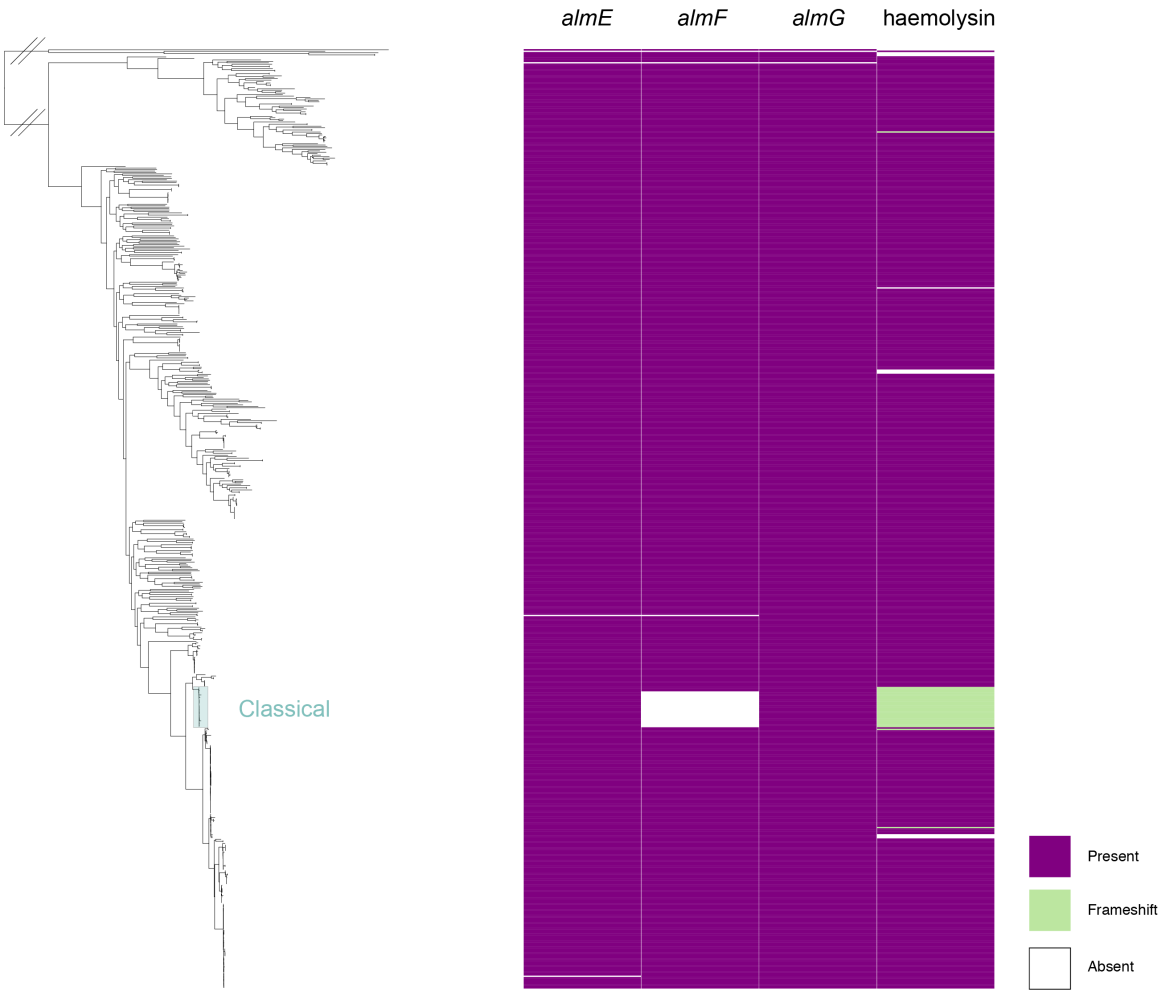


Figure 5.21 – Presence of intact and disrupted polymyxin B resistance genes and *hlyA* across the *V. cholerae* phylogeny. The absence of *almF* was determined from the pangenome matrix, as was the presence of the classical *hlyA* frameshift.

Collectively, these data show that genetic determinants which confer *V. cholerae* biotypes are, in fact, describing the Classical lineage rather than 7PET.

5.4 – Discussion

The results presented in this chapter describe some important aspects of *V. cholerae* biology, and challenge a number of dogmatic views in *V. cholerae* research. Firstly, it is not true to say that the VSP and VPI pathogenicity islands are found exclusively in epidemic or pandemic lineages of *V. cholerae*. Figures 5.5 and 5.7 show that these elements can be found in other *V. cholerae* of serogroup O1 as well as in non-O1 bacteria, consistent with previous reports (*e.g.*, [55, 396, 440]) and summarised below in Figure 5.22. Moreover, the presence of VPI-1 in non-O1 isolates underlines the fact that *V. cholerae* other than the members of pandemic lineages have the capacity to become lysogenised with CTXφ and to gain the capacity to express CT.

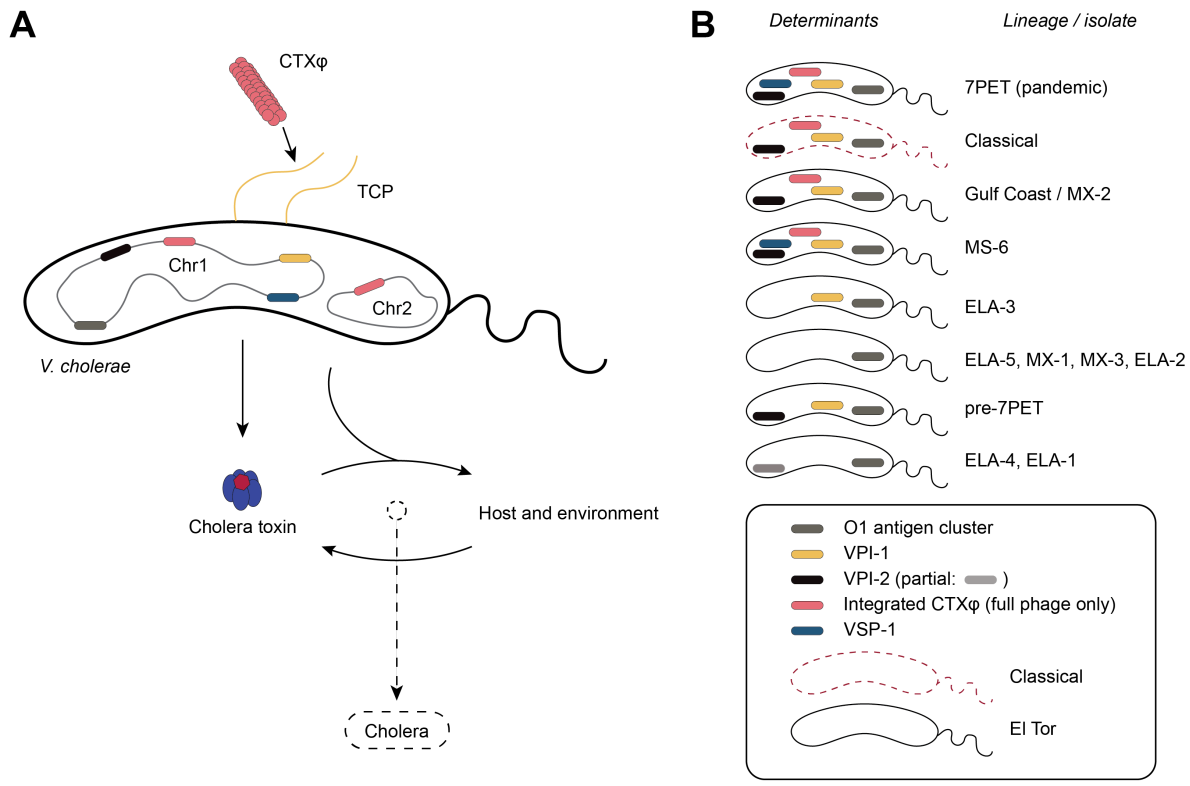


Figure 5.22 – Model of genetic determinants contributing to cholera. (A): Canonically, *V. cholerae* O1 harbouring a combination of genetic determinants can produce CT in a human host, causing cholera. However, the data presented in this chapter, as well as other work cited throughout, has shown that these genetic determinants can be found in multiple lineages of *V. cholerae* O1, not just those that cause pandemic cholera (B).

Our results suggest that the stocks of NCTC 8457 held by NCTC may not be the same as the strain that was sequenced in 2007 (Figure 5.10). The additional data available to us, including the fact that our sequenced stock is genomically of serogroup O1 and is phylogenetically

related to other non-7PET *V. cholerae* O1, and that this sequence appears to harbour a small extrachromosomal plasmid (Figures 5.13, 5.14), indicate that in order to resolve this satisfactorily, it will be necessary to obtain and re-sequence a culture of the originally-sequenced NCTC 8457 stock. Since this original sequence was generated by TIGR in the USA, it is possible that the stocks of NCTC 8457 held by ATCC (under accession # ATCC 14033; Table 5.2) and by NCTC are not identical.

To our knowledge, the data in this chapter represent the first report of a sequenced IncA/C2 plasmid which lacks any resistance determinants. Moreover, since NCTC 3661 was isolated in 1931, this suggests that pNCTC3661 is one of the oldest sequenced IncA/C2 plasmids and may be an example of an ancestral IncA/C2 plasmid backbone into which antimicrobial resistance determinants were imported as the use of antibiotics as therapeutics became more frequent. This sequence may be of considerable utility for the construction of plasmid phylogenies, or otherwise for understanding how IncA/C2 plasmids evolved to acquire AMR determinants over time. Since this plasmid type seems to be one of very few types that establish themselves in *V. cholerae* (Figures 5.1, 5.8; [158]), studying this plasmid type both within and outside the context of *Vibrio* spp. may be of considerable evolutionary interest.

This cluster of historical non-7PET serogroup O1 *V. cholerae* from NCTC underline the fact that using biotyping to identify *V. cholerae* of clinical importance is of very limited utility. These data show that all of the principal phenotypes upon which the biotyping scheme was devised describe the Classical lineage, rather than an El Tor lineage. This is exemplified by NCTC 9422 and the fact that this was co-deposited alongside other El Tor bacteria to which it was distantly related. It is apparent that the El Tor phenotype is much more broadly distributed across *V. cholerae* than just 7PET and local O1 lineages, whereas the specific mutations causing the Classical biotype appear to be coincident only in the Classical lineage. This genomic evidence is complementary to a conclusion from Chapter 3, in which the reliance on Inaba/Ogawa serotyping for epidemiological purposes was shown to be of very limited utility. Taken together, the work in this thesis so far strongly suggests that, in the light of genomic evidence, many of the bacteriological and biochemical assays used to study epidemic *V. cholerae* are misleading.

It has been observed that recent 7PET isolates, such as those which caused cholera outbreaks in Yemen, were sensitive to polymyxin B in spite of being part of an “El Tor” lineage [309]. It

was postulated that a non-synonymous mutation in *vprA* (causing a D89N mutation in VprA) was responsible for this phenotype, because functional VprA is required for the expression of the *almEFG* operon [309, 465]. This indicates that the acquisition of polymyxin B sensitivity in these recent 7PET isolates is not due to the acquisition of a Classical *almF* genotype – indeed, *almF* is intact in these recent 7PET isolates, a representative of which was included in the phylogeny discussed in this chapter (Figure 5.21). This indicates that although recent hybrid 7PET isolates appear to be acquiring "classical-like" phenotypes, this is not due to the acquisition of the same genetic mutations found in Classical isolates [221, 222, 309]. Understanding whether acquiring classical phenotypes alter the fitness of 7PET is an area of active research [466].

It appears that as diverse bacteria continue to be sequenced, the *V. cholerae* phylogeny continues to expand. Of particular interest for future study is the population of diverse genomes in the phylogeny, with a lower ANI to 7PET than the rest of the species (Figures 5.1, 5.3). It is possible that this clade corresponds to the 'novel species' or 'basal lineage' described by [467], though such a taxonomic determination would require considerable additional computational and microbiological study, and was outside the scope of this thesis research. As part of this PhD, the number of genomes available in this clade has been substantially expanded, and includes two closed genomes (NCTC 30, 48853_F01; Chapter 4). Thus, this is a promising area of future research into the diversity of the *V. cholerae* species.

In the final results chapter of this thesis, I consolidate the lines of enquiry that I have followed thus far, exploring the differences between pandemic and non-pandemic *V. cholerae*. To do this, I select eight key isolates from the phylogeny presented in this chapter, and design experiments by which to characterise the differences in the transcriptomes of these isolates when cultured under the same growth conditions. This is to explore whether gene expression patterns are more similar within or between lineages, and whether differences in regulation can be detected within lineages.