Chapter 4

Long-read sequencing of modern and historical *V. cholerae*

Contribution statement

Nick Thomson supervised the work described in this chapter. The *V. cholerae* O139 genome sequence data were made available for this PhD by Firdausi Qadri. Cultures of NCTC strains were supplied by Sarah Alexander and Julie Russell. The pACYC184 plasmid stock used for molecular cloning was a gift from Francesca Short. Jake Turnbull assisted with the collation of NCTC internal records, and Mohammed-Abbas Fazal prepared gDNA from NCTC 30 batch 4 at PHE laboratories. Leanne Kane and I prepared samples for electron microscopy, which were imaged by Claire Cormie and David Goulding. Daryl Domman generated a Canu assembly for isolate 48853_G01.

I performed all experiments and genomic analyses, and produced all figures except for microscopy images.

Publication

Some figures and data used in this chapter have been published in the following articles:

Dorman MJ, Kane L, Domman D, Turnbull JD, Cormie C, Fazal M-A, Goulding DA, Russell JE, Alexander S & Thomson NR (2019). The history, genome and biology of NCTC 30: a non-pandemic *Vibrio cholerae* isolate from World War One. *Proceedings of the Royal Society B* **286** (1900): 20182025.

Dorman MJ & Thomson NR (2020). Community evolution: Laboratory strains in the era of genomics. *Microbiology* **166** (3): 233-238. Invited "insight review" article.

Dorman MJ*, Domman D*, Uddin MI*, Sharmin S, Afrad MH, Begum YA, Qadri F & Thomson NR (2019). High quality reference genomes for toxigenic and non-toxigenic *Vibrio cholerae* serogroup O139. *Scientific Reports* **9** (1): 5865. (* Joint first author)

4.1 – Overview

In Chapter 3, I used a collection of *V. cholerae* genomes from Argentina to perform a detailed analysis of the evolution of a sub-lineage of highly clonal and genomically-invariant pandemic *V. cholerae* O1. As has been mentioned previously (section 3.5), if research efforts are focused on 7PET alone, the diversity of the species will be neglected. To that end, Chapter 3 included a high-level characterisation of genomic differences between pandemic and non-pandemic *V. cholerae*. For instance, amongst just 61 non-O1 Argentinian *V. cholerae*, examples could be seen of isolates with ANI values suggesting that they were on the boundary of being classifiable as a new species (section 3.4.10; Figure 3.28).

The fact that the diversity of the *V. cholerae* species is juxtaposed to the clonality of 7PET indicates that to understand this species further, the sequencing and characterisation of non-pandemic *V. cholerae* is required. However, although thousands of *V. cholerae* have been sequenced to date, there are still very few high-quality and accurate genome assemblies for this species, particularly for non-pandemic non-O1 *V. cholerae* isolates. High-quality genome sequences are needed to perform robust comparative genomic studies, some of which have been discussed previously (Introduction, section 1.4.1). Such comparative studies are required to address one of the most fundamental questions in *V. cholerae* research – the need to describe what discriminates pandemic and non-pandemic *V. cholerae*.

Since exploring the differences between pandemic and non-pandemic *V. cholerae* was a major aim of this PhD (section 1.5), it was decided to use long-read sequencing to generate accurate assemblies for five important *V. cholerae* isolates. These comprise four recently-isolated *V. cholerae* of serogroup O139, and a fifth historical strain isolated in 1916. Long-read technologies were used in order to resolve problems that are specific to the sequencing of *V. cholerae*, such as the presence of repeats in the integron on chromosome 2 [49, 406], and the potential for multiple copies of the CTXφ prophage to be integrated into one or both chromosomes [54, 162]. Having access to long- and short-read data for the same sample also enable hybrid assembly approaches, which use both long- and short-reads obtained from the same gDNA preparation, and allow the benefits of both PacBio RSII and Illumina technologies to be combined to produce high-quality genome assemblies [316, 319, 407], to a minimum of 'improved high-quality draft' status [407].

An overview of the history of *V. cholerae* O139 has been presented in the Introduction (section 1.3.2.1). It is known that epidemic *V. cholerae* O139 form a sub-lineage of the 7PET pandemic lineage [158, 234]. Although a serogroup O139 isolate obtained from a patient in India during 1992, dubbed MO10 [241], has been sequenced and used for comparative genomics in the past [54, 133], a closed genome for this clinically- and epidemiologically-important 7PET sub-lineage had not been reported prior to this PhD work. In this project, PacBio sequences for three toxigenic and one non-toxigenic *V. cholerae* O139 were assembled and analysed [244]. All four isolates were of clinical origin [244].

V. cholerae is a pathogen controlled under Schedule 5 of the Anti-Terrorism, Crime and Security Act (ATCSA) in the United Kingdom [408]. This means this bacterium must be worked on in a laboratory that meets specific security requirements. The only facilities at WSI that meet these requirements are a subset of our Containment Level 3 (CL3) laboratories. Therefore, as part of establishing a CL3 laboratory at WSI for conducting research into *V. cholerae*, a reliable protocol for the efficient extraction of high molecular weight gDNA from this species was required (Methods, section 2.2.5). Large fragments of gDNA that have not been mechanically sheared are required for long-read sequencing. The first *V. cholerae* isolate chosen to be cultured for PacBio sequencing using this methodology was NCTC 30, the oldest *V. cholerae* accessioned by NCTC, and to our knowledge, the oldest live isolate of this species that is publicly available for research. NCTC 30 is a unique scientific curiosity, as well as an important historical *V. cholerae* isolate.

4.2 – Specific aims

The work described in this chapter aimed to

1) Assemble, annotate, and characterise *V. cholerae* O139 genomes sequenced with long-read technologies, to generate reference assemblies for this medically-important sub-lineage of 7PET,
2) Sequence to completion the genome of NCTC 30, an historical curiosity and an important example of a non-O1/O139 *V. cholerae* which caused 'choleraic diarrhoea', and
3) Capitalise on having access to a live culture of NCTC 30 to validate genomic observations experimentally, and to optimise experimental protocols for use at CL3.

4.3 – Results

4.3.1 – Closed genome assemblies for toxigenic and non-toxigenic V. cholerae *O139*

Four previously-reported *V. cholerae* of serogroup O139 [244] had been re-sequenced using PacBio RSII technology prior to the beginning of this project, in order to generate reference-grade assemblies for genomic analyses. Illumina short-reads were also available for these isolates [244]. *De novo* assemblies of these PacBio reads were generated and annotated as described in Chapter 2.

Using the long-reads for each of these genomes, single contig assemblies were produced for each of the two chromosomes in each isolate, which were then corrected and circularised (Methods, section 2.1.2.2). The short-read data for each isolate were used to correct the assemblies, but this correction step did not make any improvements to the PacBio-only assemblies. Accordingly, it was decided to proceed with using these long-read assemblies for subsequent analyses in this chapter. Coverage percentages and other summary statistics for these assemblies are listed in Table 4.1.

Internal sequence ID	**Sample name**	**CTXφ present?**	**Genome size (bp)**	**Coverage of *de novo* assembly with long reads (%)**	**Coverage of N16961 (%)**	**SNVs relative to N16961**
48853_F01	MP_070116	No	4123525	165.76	58.5	122865
48853_G01	P_0684000	Yes	4092641	170.56	97	271
48853_H01	ICVB_2236_02	Yes	4092645	147.29	97	270
48853_A02	SMIC_67_01	Yes	4092644	165.65	97	274

Table 4.1 - Summary statistics for four closed *V. cholerae* O139 assemblies. Potentially recombined SNVs were not excluded from this analysis because 48853_F01 was extremely genetically distant from N16961. Modified from [409]. The HGAP assembler assembled the reads from sample 48853_G01 into three contigs. Re-assembling this sample with Canu v1.1 [410] produced a two-contig assembly, which was used for subsequent analysis.

4.3.2 – Phylogenetic position of toxigenic V. cholerae *O139*

The phylogenetic position of the three toxigenic sequences was confirmed by placing these isolates into context with additional *V. cholerae* O139 sequences [242, 244] and other 7PET

genomes [234]. A maximum-likelihood phylogeny was calculated from an alignment of 1,630 non-recombinant SNVs, which had been identified by mapping short-read data to the N16961 reference sequence (Figure 4.1). Since the three toxigenic isolates varied in length by four bases at most (Table 4.1), and were phylogenetically identical to one another, just one of the assemblies (48853_H01) was chosen to be used as an exemplar sequence for all subsequent comparative analyses.

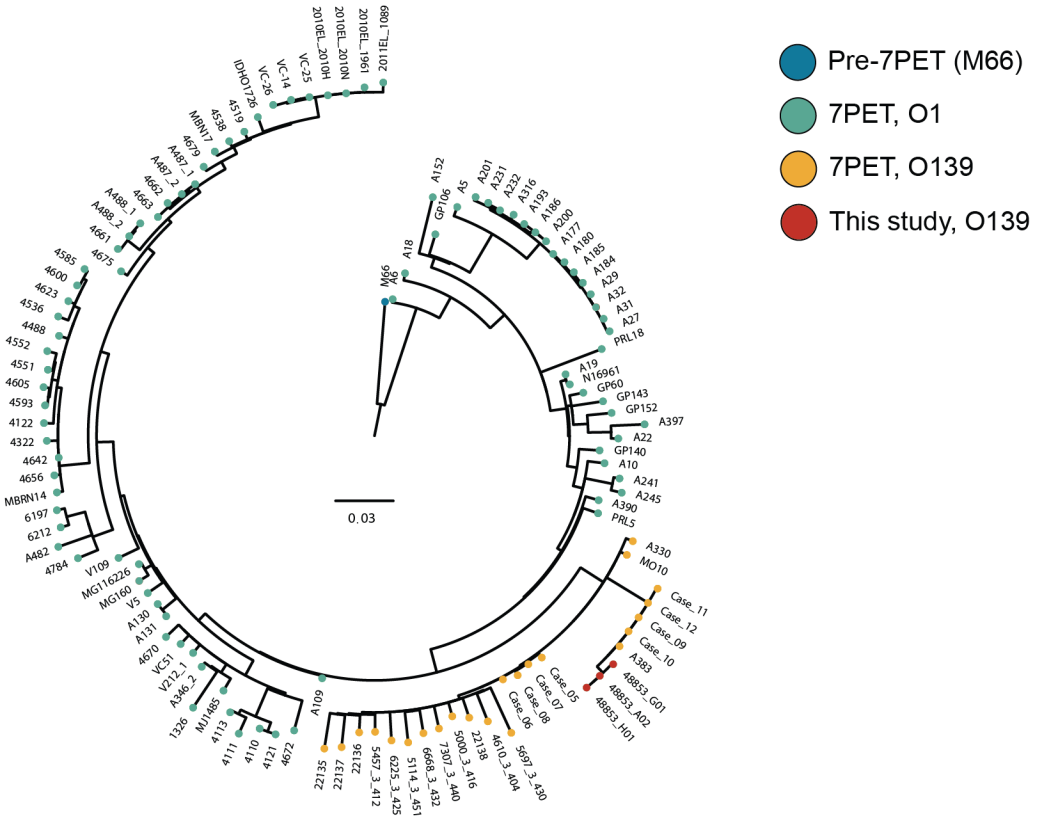

Figure 4.1 – Maximum-likelihood phylogenetic tree of 7PET. Phylogeny rooted on the M66 outgroup sequence. Scale bar denotes the number of substitutions *per* variable site. This phylogeny was computed under the GTR+Gamma model using RAxML v8.2.8 with 500 bootstrap replicates [411]. Modified from [409].

4.3.3 –CTXφ prophage sequences in toxigenic V. cholerae *O139*

Previous work had reported that multiple types and arrangements of CTXφ prophage had been detectable in *V. cholerae* O139 chromosomes, depending on the year and location from which an isolate was isolated [230]. In particular, isolates obtained from Calcultta in 1996 had been reported to harbour two types of prophage, CTXφ^{cal} and CTXφ$^{El\ Tor}$ in the configuration CTXφ$^{El\ Tor}$ - CTXφ^{cal} - CTXφ^{cal} [412–414]. However, the assembly for MO10, used by many as an O139 reference isolate [241] was insufficiently well-assembled to be confident of the

context of CTXφ in this laboratory strain; this assembly is currently available in 84 contigs (accession # GCA_000152425.1). Similarly, the short-read data from the three more recently sequenced *V. cholerae* O139 isolates could not be assembled across the CTXφ region into a single contig, and in those assemblies, only one of the two *ctxB* genes was identifiable (the Illumina assemblies for 48853_G01 and 48853_H01 contained only a *ctxB4* allele in a small contig, and 48853_A02 contained *ctxB5* in a larger contig).

Using the long-read assemblies generated here, it was evident that 48853_H01 contains three CTXφ prophages arranged in tandem, at the same integration site on the larger chromosome of N16961, between the *VC_1450* and *VC_1467* loci (Figure 4.2). This was the case in all three toxigenic *V. cholerae* O139 sequenced here (data not shown). In order to verify the presence of multiple copies of CTXφ in these genomes, the Illumina reads previously reported for this isolate were mapped to the N16961 reference and the coverage of these mapping data over the CTXφ region were visualised (Figure 4.3). The CTXφ region was covered 2-3 times as highly as the surrounding chromosome, supporting further the existence of multiple CTXφ copies in these *V. cholerae* O139 genomes.

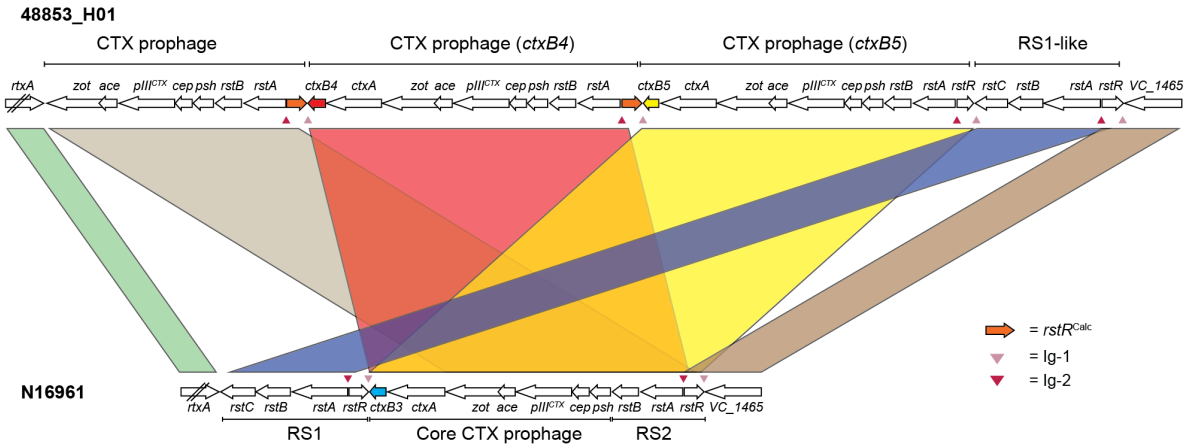

Figure 4.2 – Comparison of the CTXφ region in assembly 48853_H01 and N16961. Region annotated as *per* [99]. Reproduced from [409].

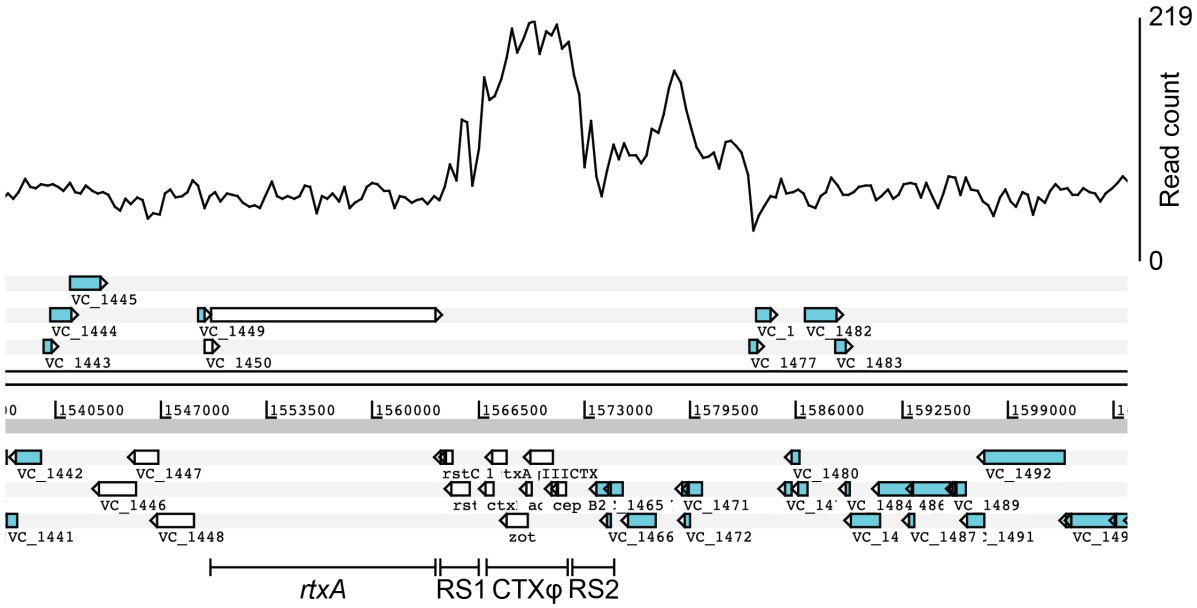

Figure 4.3 – Validating the presence of multiple CTXφ copies by mapping. Figure depicts the depth of short-reads for *V. cholerae* O139 isolate 48853_H01 mapped to the N16961 reference sequence. Mapping coverage was plotted using Bamview and Artemis [339, 356]. Annotation as described in Figure 4.2. Reproduced from [409].

Two of these repeats harboured distinct *ctxB* alleles, *ctxB4* and *ctxB5*. The *ctxB* gene closest to *rtxA* in these assemblies was a *ctxB5* allele, and the second *ctxB* was a *ctxB4* allele. Both *ctxB* alleles have been found in *V. cholerae* O139 strains previously [415, 416]. The third CTXφ repeat was partial, and lacked the *ctxAB* operon while comprising the genes between and including *zot* and *rstA*, and an *rstR* open reading frame corresponding to $rstR^{Calc}$ [417] (Figure 4.2). A complete *attL* sequence was identified adjacent to the *VC_1465* locus in 48853_H01 [96]. The phage sequence in the *attR* site adjacent to *rtxA* is not identical to that reported by Huber and Waldor [96], although the *attR* element does contain the central recombination motif and the residual bacterial *attB* sequence (see section 1.2.3 for an overview of CTXφ integration mechanisms). Although it has been reported that *V. cholerae* O139 can harbour more than one type of CTXφ phage simultaneously [230, 413, 414, 417], we were unaware of previous reports of multiple *ctxB* alleles co-existing in the same *V. cholerae* genome.

The *ctxB4* and *ctxB5* alleles differ from one another at two nucleotide positions; 83 and 115 (Figure 4.4A). These produce corresponding non-synonymous substitutions at amino acids 28 and 39 (Figure 4.4B). Additional confirmation that these alleles co-exist in the *V. cholerae* O139 genomes was obtained by inspecting manually the sequences of reads mapped to this

locus (Figure 4.5), and from the fact that the assemblies for all three toxigenic *V. cholerae* O139 harboured both *ctxB* alleles, in CTXφ that were in the same chromosomal arrangement in all three genomes (Figure 4.2). It was also confirmed by manual inspection of the assemblies that no CTXφ prophage were present in the smaller chromosome in these isolates.

A

```
                1
ctxB1       ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC
ctxB3/B3b   ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC
ctxB4       ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC
ctxB5       ATGATTAAAT TAAAATTTGG TGTTTTTTTT ACAGTTTTAC TATCTTCAGC ATATGCACAT GGAACACCTC

               71
ctxB1       AAAATATTAC TGATTTGTGT GCAGAATACC ACAACACACA AATACATACG CTAAATGATA AGATATTTTC
ctxB3/B3b   AAAATATTAC TGATTTGTGT GCAGAATACC ACAACACACA AATATATACG CTAAATGATA AGATATTTTC
ctxB4       AAAATATTAC TGATTTGTGT GCAGAATACC ACAACACACA AATATATACG CTAAATGATA AGATATTTTC
ctxB5       AAAATATTAC TGCTTTGTGT GCAGAATACC ACAACACACA AATACATACG CTAAATGATA AGATATTTTC

              141
ctxB1       GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AACTTTTCAA
ctxB3/B3b   GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AATTTTTCAA
ctxB4       GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AACTTTTCAA
ctxB5       GTATACAGAA TCTCTAGCTG GAAAAAGAGA GATGGCTATC ATTACTTTTA AGAATGGTGC AACTTTTCAA

              211
ctxB1       GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA
ctxB3/B3b   GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA
ctxB4       GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA
ctxB5       GTAGAAGTAC CAGGTAGTCA ACATATAGAT TCACAAAAAA AAGCGATTGA AAGGATGAAG GATACCCTGA

              281
ctxB1       GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT
ctxB3/B3b   GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT
ctxB4       GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT
ctxB5       GGATTGCATA TCTTACTGAA GCTAAAGTCG AAAAGTTATG TGTATGGAAT AATAAAACGC CTCATGCGAT

              351
ctxB1       TGCCGCAATT AGTATGGCAA ATTAA
ctxB3/B3b   TGCCGCAATT AGTATGGCAA ATTAA
ctxB4       TGCCGCAATT AGTATGGCAA ATTAA
ctxB5       TGCCGCAATT AGTATGGCAA ATTAA
```

B

Signal sequence (21 AA)

```
                1
CtxB1       MIKLKFGVFF TVLLSSAYAH GTPQNITDLC AEYHNTQIHT LNDKIFSYTE SLAGKREMAI ITFKNGATFQ
CtxB3/B3b   MIKLKFGVFF TVLLSSAYAH GTPQNITDLC AEYHNTQIYT LNDKIFSYTE SLAGKREMAI ITFKNGAIFQ
CtxB4       MIKLKFGVFF TVLLSSAYAH GTPQNITDLC AEYHNTQIYT LNDKIFSYTE SLAGKREMAI ITFKNGATFQ
CtxB5       MIKLKFGVFF TVLLSSAYAH GTPQNITALC AEYHNTQIHT LNDKIFSYTE SLAGKREMAI ITFKNGATFQ

               71
CtxB1       VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
CtxB3/B3b   VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
CtxB4       VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
CtxB5       VEVPGSQHID SQKKAIERMK DTLRIAYLTE AKVEKLCVWN NKTPHAIAAI SMAN*
```

Figure 4.4 – Alignment of *ctxB* and CtxB variants. Highlighted *ctxB* allelic variations (A) give rise to non-synonymous variations in encoded CtxB proteins (B). The *ctxB1* allele is associated with Classical *V. cholerae*, and *ctxB3* with wave 1 7PET *V. cholerae* such as N16961 [59, 97, 234]. None of these mutations are located in the region of *ctxB* which encodes the N-terminal signal sequence (annotation taken from [77]).

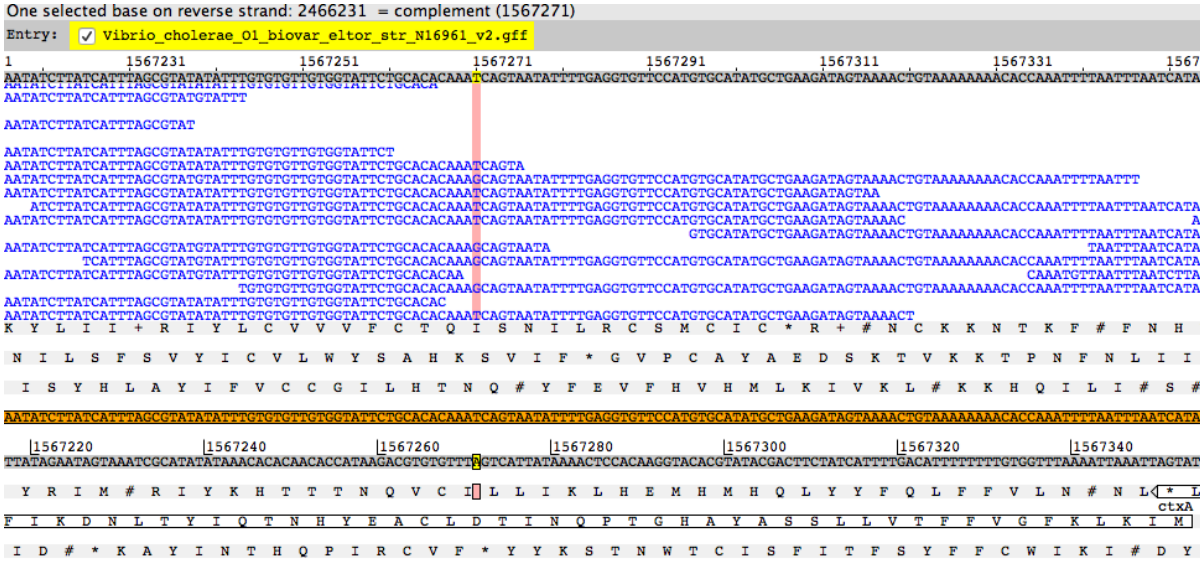

Figure 4.5 – Reads corresponding to both *ctxB4* and *ctxB5* from toxigenic *V. cholerae* O139 map to N16961. In this example, the base at position 83 in *ctxB* has been highlighted (see Figure 4.4). These reads are mapped to the opposite strand to the annotated *ctxB*.

Since multiple CTXφ configurations have been reported in *V. cholerae* O139, and in order to broaden our observations, a set of 50 *V. cholerae* O139 genomes that were obtained during a study of intra-household cholera transmission in Dhaka, Bangladesh were re-examined [235]. These formed two sub-lineages within 7PET (dubbed B3 and B4) and were isolated in 2002 and 2003, pre-dating the isolates sequenced in this study. Forty-one of these isolates harboured both *ctxB4* and *ctxB5* alleles simultaneously, all of which were in lineage B4. The short-read data for the genomes sequenced in this chapter had been included in the Dhaka household transmission study, and were closely related to B4 [235]. The remaining 9 only harboured one *ctxB* allele, *ctxB4*. All of these were in lineage B3. This suggests that there is genomic complexity and population structure within the O139 sub-lineage of 7PET that has been previously-hidden due to the lack of an accurate reference sequence for the sub-lineage.

4.3.4 – Genomic island complements of V. cholerae *O139 isolates*

Having seen variation in CTXφ, the complement of other genomic islands contained in the three toxigenic and one non-toxigenic *V. cholerae* O139 isolates' genomes was determined. Examined were the canonical pathogenicity islands, prophages, the SXT ICE*Vch*Ind4 element, and other genomic islands that have been reported previously in *V. cholerae* [54, 55, 133, 134, 141]. VPI-1 and VSP-2 were present in all three toxigenic *V. cholerae* O139 isolates (Table

4.2; Figure 4.6). Moreover, VPI-2 was severely truncated to the point of absence in these three isolates. This truncation has been described previously in MO10 and is a hallmark of toxigenic *V. cholerae* O139 [54, 418] (Table 4.2; Figure 4.6). VPI-1 was partially present in the genome of the non-toxigenic isolate (Table 4.2).

Sample Name	**VSP-1 (*VC_0174-VC_0186*)**	**VSP-2 (*VC_0489-VC_0517*)**	**VPI-1 (*VC_0809-VC_0848*)**	**VPI-2 (*VC_1757-VC_1810*)**	**CTXφ (*VC_1451-VC_1465*)**	**SXT (*VC_0659* insertion)**
48853_F01	Absent	Absent	Partially present (deletion of *VC_0817-VC_0848*)	Absent	Absent	64% match to ICE*Vch*Ind4
48853_G01	Present, and duplication on chr2	Present	Present	Deletion of *VC_1761-1787*	Present, in more than one copy	100% match to ICE*Vch*Ind4
48853_H01	Present, and duplication on chr2	Present	Present	Deletion of *VC_1761-1787*	Present, in more than one copy	100% match to ICE*Vch*Ind4
48853_A02	Present, and duplication on chr2	Present	Present	Deletion of *VC_1761-1787*	Present, in more than one copy	100% match to ICE*Vch*Ind4

Table 4.2 – Presence and absence of select genomic islands in *V. cholerae* O139 genome assemblies. Similarity percentages were obtained by comparing SXT element sequences to that of ICE*Vch*Ind4 using BLASTn. chr2 = chromosome 2. Modified from [409].

A genomic island integrated into the *VC_0659* locus (encoding peptide chain release factor 3) was detected in each of the three toxigenic *V. cholerae* O139 assemblies. This island was identical to SXT (also known as ICE*Vch*Ind4 [156]) and was integrated into the same locus as is SXT in MO10 [156]. An insertion into *VC_0659* was also identified in the non-toxigenic O139 genome assembly, which was 64% identical to ICE*Vch*Ind4 (Table 4.2). These observations are fully consistent with previous data on genomic island distribution in the MO10 genome [54].

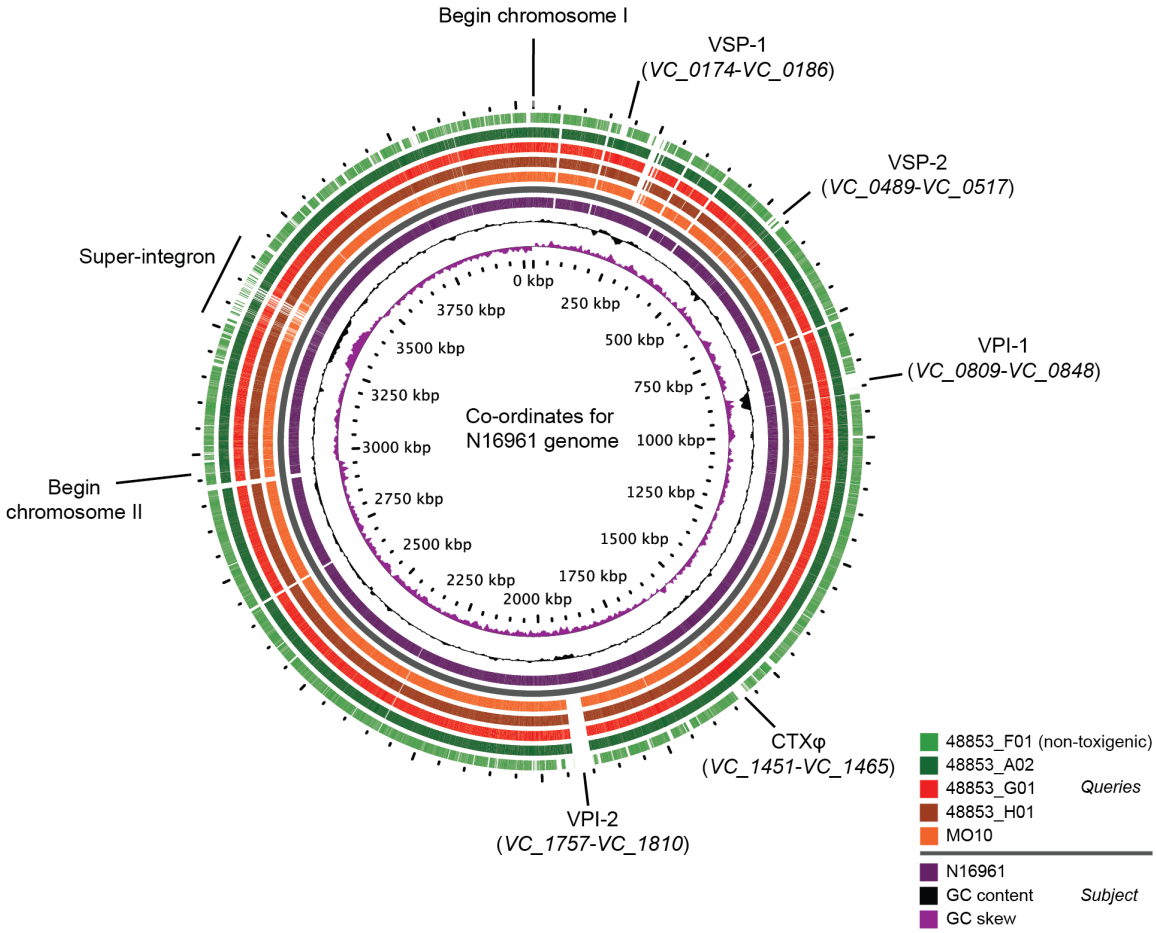

Figure 4.6 – BLAST atlas illustrating the location, presence, and absence of key genomic islands in *V. cholerae* O139 assemblies relative to the N16961 reference sequence. The sequences of both N16961 chromosomes were concatenated to produce this figure. Reproduced from [409].

VSP-1 was integrated on the larger chromosome between genes *VC_0173* and *VC_0187* in the three toxigenic *V. cholerae* O139 isolates, as it is in N16961 [54] (Figure 4.6; Table 4.2). However, a 14.3 kb sequence of DNA was also detected on the smaller chromosome of each of the toxigenic isolates, integrated between *VC_A0695* and *VC_A0696*, that was 99% identical at the nucleotide level to VSP-1 (*VC_0175* to *VC_0186*; Figure 4.7A). This strongly suggested that a second copy of the VSP-1 element was present on the second chromosome in each of these genomes. In order to verify this, the short-reads for each isolate were mapped to the N16961 reference genome the read depth was plotted for VSP-1 relative to the surrounding genome (Figure 4.7B).

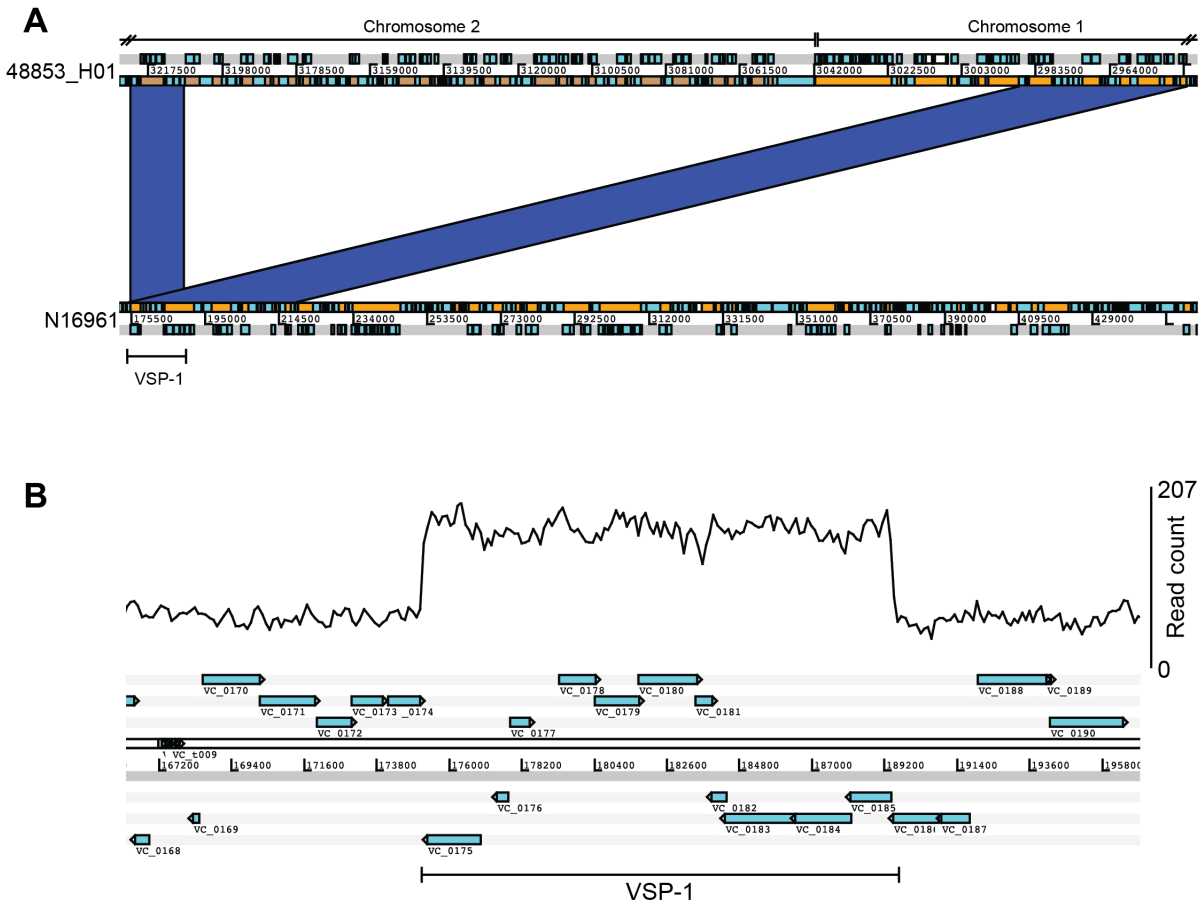

Figure 4.7 – Presence of VSP-1 on both chromosomes of *V. cholerae* O139. DNA that was homologous to VSP-1 was detected on both chromosomes of 48853_H01 (A). Mapping short-reads from 48853_H01 to N16961 confirmed that there were approximately double the number of sequencing reads for VSP-1 relative to the surrounding chromosome (B). Modified from [409].

This VSP-1 duplication was detected in each of the three toxigenic *V. cholerae* O139 genome assemblies (Table 4.2). It is known that VSP-1 is capable of excising from the larger *V. cholerae* chromosome [141], and it has been previously reported that the *V. cholerae* O1 isolate MJ-1236 (the Matlab variant) harbours a second copy of VSP-1 integrated between *VC_A0695* and *VC_A0696* [419]. Separately, Grim *et al*. identified a single clinical isolate of *V. cholerae* O139 from Bangladesh that appeared to harbour an insertion between *VC_A0695* and *VC_A0696* that resembled VSP-1, but this isolate was not described further [419]. This is likely to be the same phenomenon observed and confirmed to be present in these three *V. cholerae* O139 genomes.

4.3.5 – Antimicrobial resistance and accessory virulence determinants

Since the ambition of this project was to use these assemblies as reference sequences for future studies of *V. cholerae* O139, having made comparisons to N16961, these assemblies were then compared to the MO10 sequence. This allows for consistency amongst these data and comparisons with previous publications (Figure 4.8). MO10 harbours the VSK prophage GI-16 [54], which N16961 lacks. GI-16 is also absent from the four sequences in this study (Figure 4.8). Likewise, MO10 harbours a kappa prophage (genomic island GI-11 [54]), which is neither present in N16961 nor in the O139 sequences characterised here (Figure 4.8). Importantly, MO10 does not appear to harbour the second VSP-1 copy on chromosome 2 identified in these three isolates, which were isolated more recently than MO10.

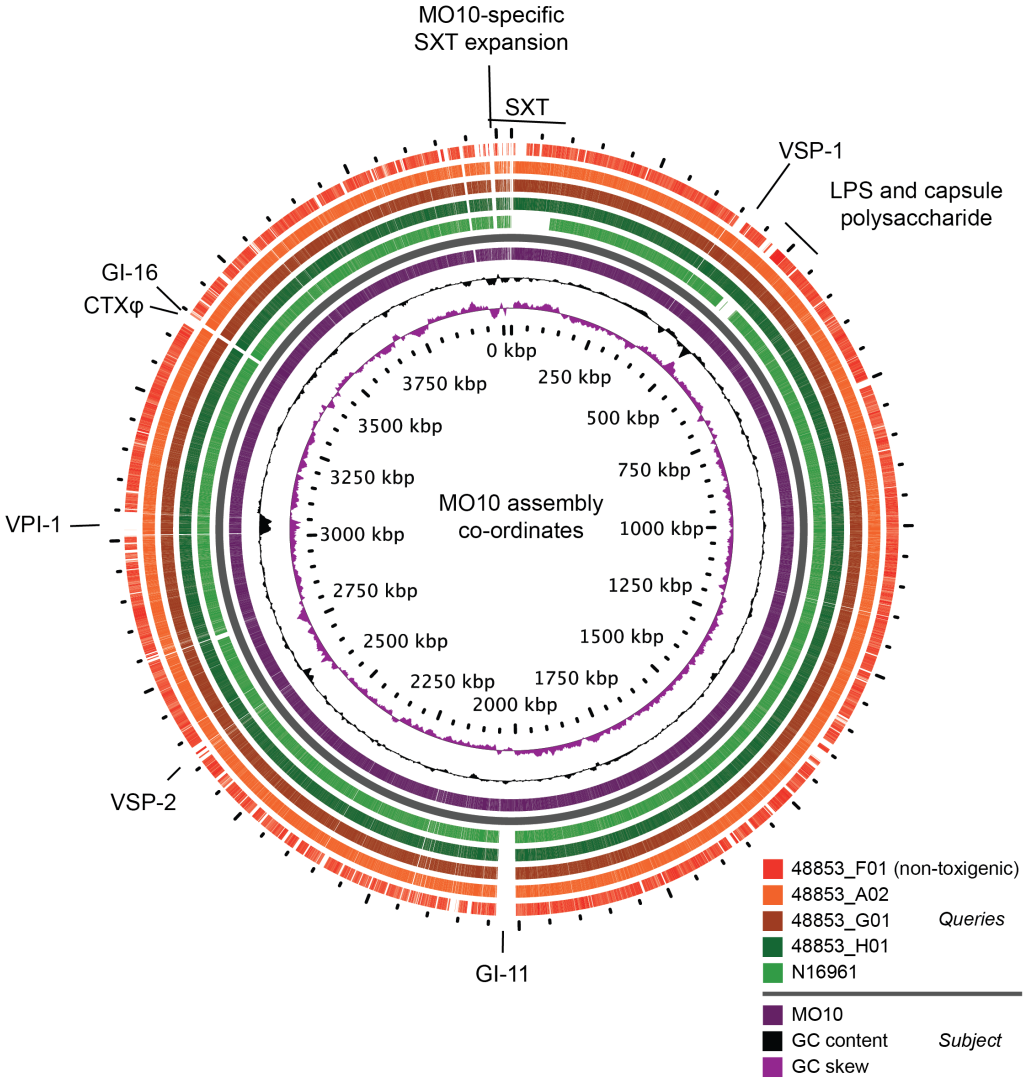

Figure 4.8 – BLAST atlas comparing *V. cholerae* O139 assemblies to MO10. The absence of O139-specific LPS and capsule genes from N16961 is evident, as is the absence of pathogenicity islands and CTXφ from the non-toxigenic isolate. All MO10 contigs were concatenated to produce this figure. Reproduced from [409].

MO10 harbours an SXT element [156] which is expanded relative to that found in these genomes (Figure 4.8). The MO10-specific SXT expansion includes genes conferring resistance to the antimicrobials streptomycin (*strAB*), sulfamethoxazole (*sul2*), trimethoprim (*dfr18*), and chloramphenicol (*floR*). Accordingly, the assemblies for all four *V. cholerae* O139 genomes were scanned for antimicrobial resistance genes. No antimicrobial resistance genes were detected in the non-toxigenic 48853_F01 genome. The three toxigenic isolate genomes also do not contain antimicrobial resistance genes with the exception of one *catB9* gene that is common to almost all 7PET *V. cholerae* [158] (Chapter 3, Figure 3.18), and is known not to render isolates resistant to phenicols [420]. These results agree with the original antimicrobial sensitivity testing of these isolates, which found that they were resistant only to nalidixic acid [244]. Examination of the *gyrA* and *parC* genes in these sequences confirmed that these four isolates harbour mutations predicted to result in an S83I mutation in GyrA. The non-toxigenic isolate 48853_F01 contains an additional mutation in *gyrA* (predicted to cause substitution A171S) and a S85L mutation in *parC* (predicted to cause an S85L substitution in ParC). All of these mutations are associated with nalidixic acid resistance in *V. cholerae* [155].

The four genome assemblies were also scanned for the presence of *V. cholerae* accessory virulence genes, with the specific objective of determining whether candidate virulence genes were present in the genome of the otherwise non-toxigenic *V. cholerae* O139 isolate, given that this isolate was obtained from a patient suffering from severe diarrhoea [244]. However, no known virulence determinants were detected in the 48853_F01 genome assembly other than those typically found across the *V. cholerae* species [1] (Table 4.3).

Accessory virulence gene	Present in 48853_F01	Present in 48853_G01	Present in 48853_H01	Present in 48853_A02
ToxR (*VC_0984*)	Yes	Yes	Yes	Yes
Zona occludens toxin, Zot (*VC_1458*)	No	Yes	Yes	Yes
Accessory cholera enterotoxin, Ace (*VC_1459*)	No	Yes	Yes	Yes
Haemolysin, *hlyA* (*VC_A0219*)	Yes	Yes	Yes	Yes
Mannose-sensitive haemagglutinin, MSHA (*VC_0398..VC_0414*)	Yes	Yes	Yes	Yes
MARTX toxin, *rtxA* (*VC_1451*)	Yes	Yes	Yes	Yes
MARTX toxin accessory gene, *rtxC* (*VC_1450*)	Yes	Yes	Yes	Yes
HA/protease, *hapA* (*VC_A0865*)	Yes	Yes	Yes	Yes
Heat-stable enterotoxin NAG-ST (Accession # M85198.1)	No	No	No	No
Type III secretion system from *V. cholerae* AM_19226 (typically present *in lieu* of VPI-2; accession # AATY01000000)	No	No	No	No

Table 4.3 – Accessory virulence genes in *V. cholerae* O139. Locus IDs from the N16961 reference, or accession numbers for sequences absent from N16961, are provided. Reproduced from [409].

4.3.6 – Phylogenetic position of non-toxigenic V. cholerae *O139*

It had been reported previously that isolate 48853_F01 was phylogenetically distantly related to the three toxigenic isolates [244]. In order to contextualise this isolate, the genomes of 61 additional *V. cholerae*, and those of three non-cholera Vibrios, were used to calculate an alignment of 2,103 core genes. From this sequence alignment, 168,476 SNVs were identified, which were used to calculate a core-gene phylogeny for these isolates (Figure 4.9).

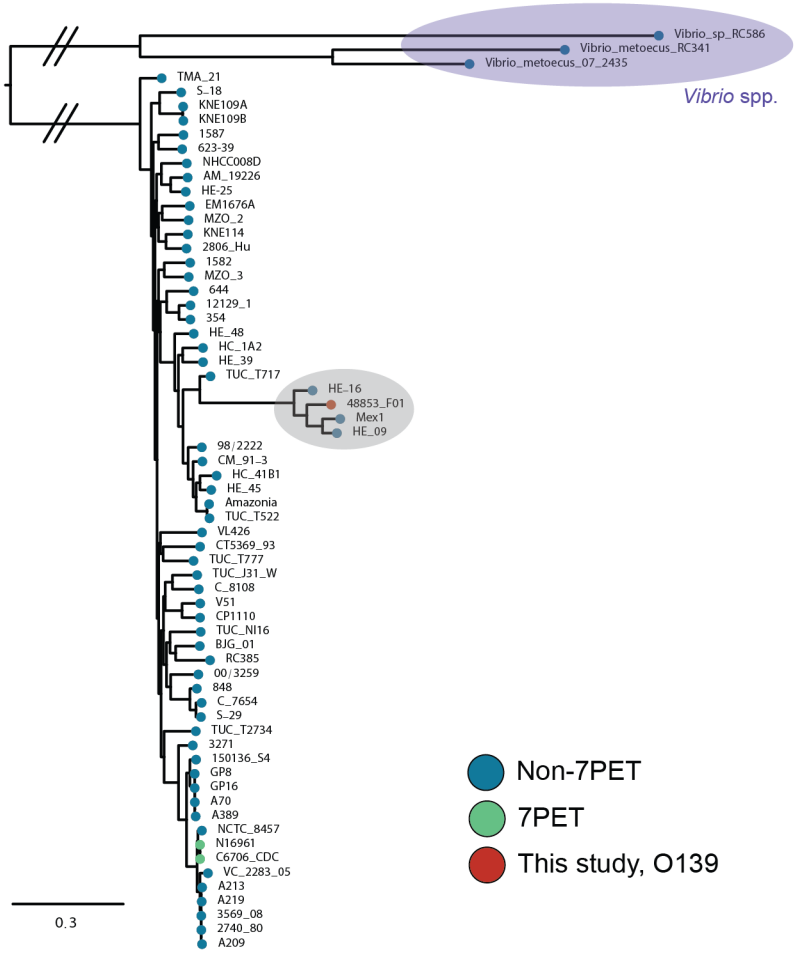

Figure 4.9 – A phylogeny of non-7PET *V. cholerae*. Rooted on three *Vibrio* spp. genomes. Hatch marks denote branches which have been manually shortened for illustrative purposes. This phylogeny was computed under the GTR+Gamma model using RAxML v8.2.8 with 500 bootstrap replicates [411]. Modified from [409].

The phylogenetic analysis showed that 48853_F01 forms a cluster with three isolates that are distantly related to the three toxigenic *V. cholerae* O139, falling outside of the 7PET lineage to which they belong (Figures 4.1, 4.9; grey disc). This cluster includes two Haitian non-O1 *V. cholerae* isolates from 2010 and a Mexican isolate from 1991 – unfortunately, serotype data were not recorded for these isolates (Figure 4.9). Moreover, several Argentinian *V. cholerae* sequenced in Chapter 3 were found to cluster with these sequences (shown in Figure 3.21).

Given the phylogenetic position of this non-toxigenic isolate, the capsule and lipopolysaccharide (LPS) biosynthesis loci in this genome were compared to the MO10 sequence and to those *V. cholerae* O139 that belonged to the 7PET lineage. Using closed genome assemblies generated here, it was confirmed that the three toxigenic strains contain O139 LPS operons that strongly resemble that found in MO10 (Figure 4.10). The equivalent

region in the non-toxigenic isolate 48853_F01 is less similar, although this strain exhibits a strong O139-positive phenotype using the rapid dipstick assay and slide agglutination tests [244]. It was also noted that the three non-O1 Haitian and Mexican *V. cholerae* which cluster with the non-toxigenic O139 isolate share capsule biosynthesis genes with 48853_F01, but they do not harbour the same LPS operon (Figures 4.9, 4.10). These isolates therefore are unlikely to be *V. cholerae* of serogroup O139.

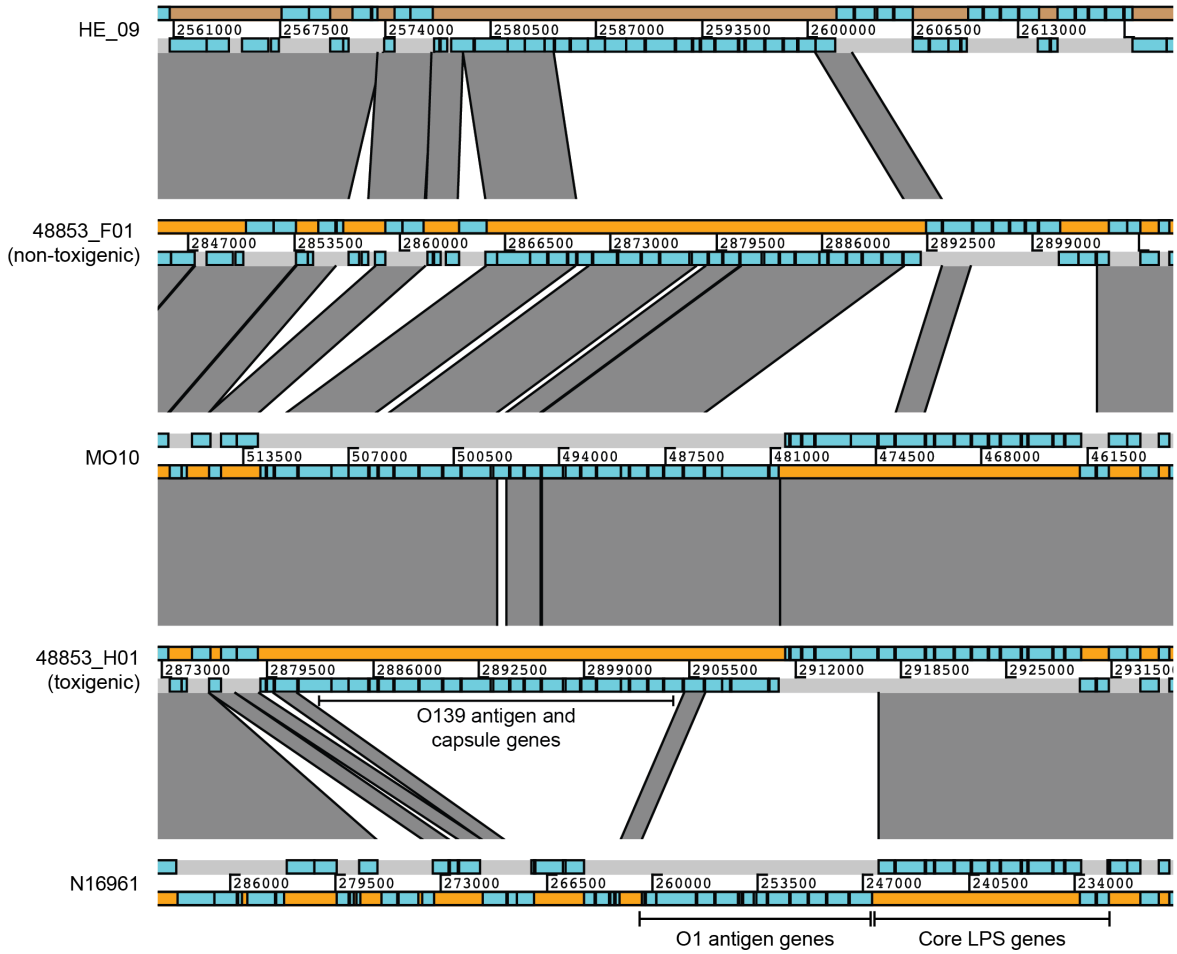

Figure 4.10 – O-antigen chromosomal loci in select *V. cholerae*. The loci known to encode the O1 and O139 antigens, as well as capsule biosynthesis genes, are indicated. HE_09 is phylogenetically related to the non-toxigenic *V. cholerae* O139 (Figure 4.9) but lacks the O139 biosynthesis operon. Reproduced from [409].

Serendipitously, the phylogenetic position of the non-toxigenic isolate 48853_F01 was made even more intriguing in the context of work on NCTC 30, the genomic and experimental characterisation of which is now presented.

4.3.7 – NCTC 30 genome sequencing and assembly

NCTC 30 was isolated in 1916 and was accessioned into NCTC in 1920 (Figure 4.11). This *V. cholerae* was originally named "Martin 1", and was isolated from a British soldier convalescent in Egypt against the background of World War 1 (WW1). This was a period in history during which cholera's epidemic potential was both recognised and feared [421, 422]. In spite of this, the British Expeditionary Forces remained largely free of cholera throughout this period. It has been estimated that the British Expeditionary Forces incurred 11,096,338 casualties during WW1, but reportedly experienced just 1,918 cholera cases in 1916, 209 cases in 1917, and 450 cases in 1918 [423]. Of these cholera patients, 49 died in 1917, and 106 died in 1918 [423].

NCTC 30 has been described as being 'probably' serogroup O2, and its biochemical profile has been described as part of its maintenance as the thirtieth strain to be accessioned into the NCTC culture collection (Figure 4.11). NCTC 30 stocks have been maintained since the strain was accessioned into the collection (Figure 4.12).

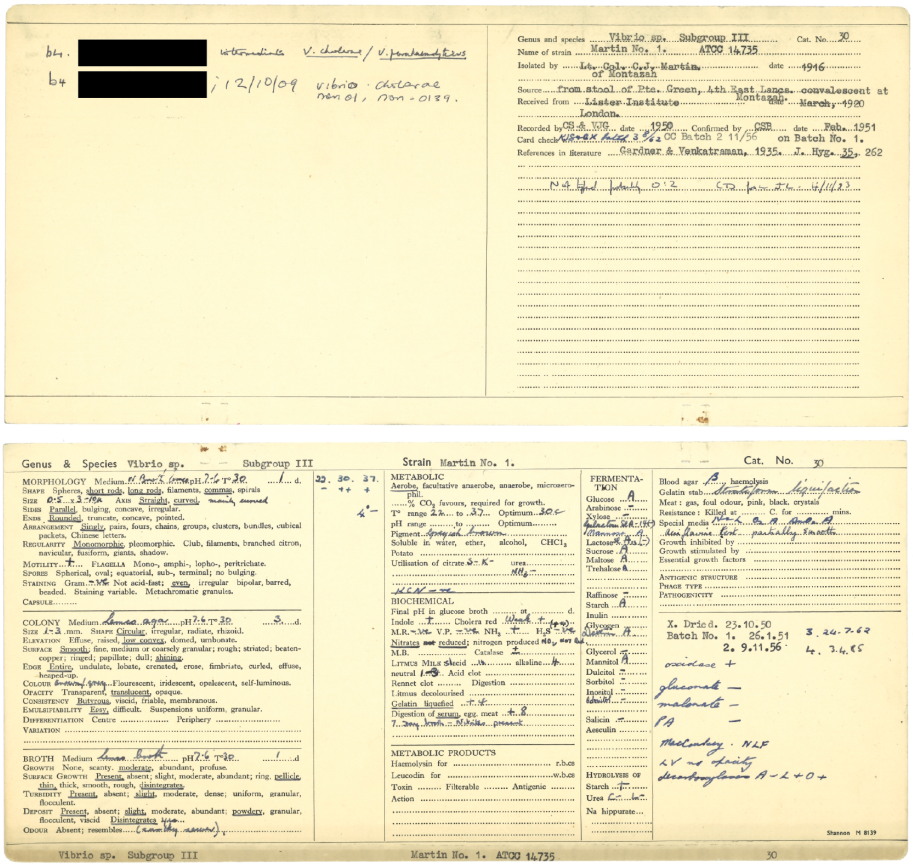

Figure 4.11 – The NCTC 30 check card. The card details the provenance of the isolate as well as aspects of its original biochemical and microbiological characterisation. Reproduced from supplementary material of [424].

For this PhD, a freeze-dried culture of NCTC 30 was revived and sequenced to completion using both long- and short-read technologies. NCTC has produced four lyophilised batches of NCTC 30 to date (Figure 4.11, 4.12), and the culture sequenced in this study was a derived from the earliest batch of lyophilised bacteria available for research (batch 3, lyophilised in 1962; Figures 4.11; 4.12). This was to minimise the risk of sequencing a culture that had acquired mutations as a consequence of long-term passage or maintenance. Molecular and genetic experiments were then carried out, to characterise further the biology of this historical isolate, and to validate genomic results.

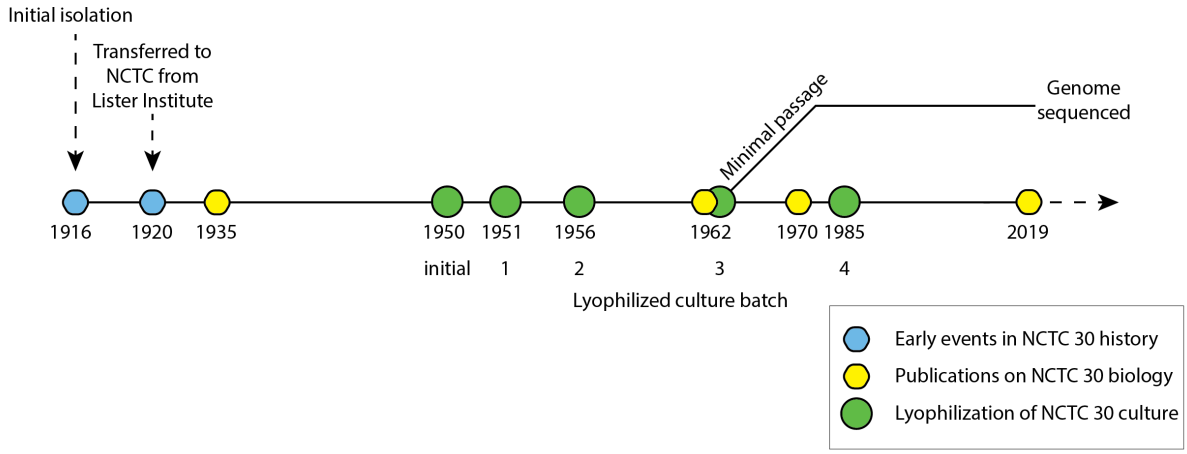

Figure 4.12 – A timeline of events in the history of NCTC 30. Figure based on data contained in Figure 4.11 and reproduced from [425]. Publications on NCTC 30 that are discussed throughout this chapter have been listed (not exhaustive).

gDNA extracted from one clone of minimally-passaged NCTC 30 (MJD382) was sequenced using both short- and long-read technologies (Illumina HiSeq X10 and PacBio RSII). The long-read assembly strategy used for these data (see Methods) produced two circular contigs, and the long reads covered the finished assembly to a depth of 148.01 X. The NCTC 30 genome assembly consisted of two circularised contigs, one corresponding to the larger chromosome of 2,922,904 bases, and one to the smaller chromosome of 1,029,451 bases (Figure 4.13). To guard against genetic rearrangements or contamination introduced through long term storage, a technical replicate extraction of gDNA from MJD382, as well as a biological replicate (gDNA from MJD439, a second colony of NCTC 30 obtained from the first passage of NCTC 30 after rehydration) were also sequenced using both long- and short-read sequencing technologies.

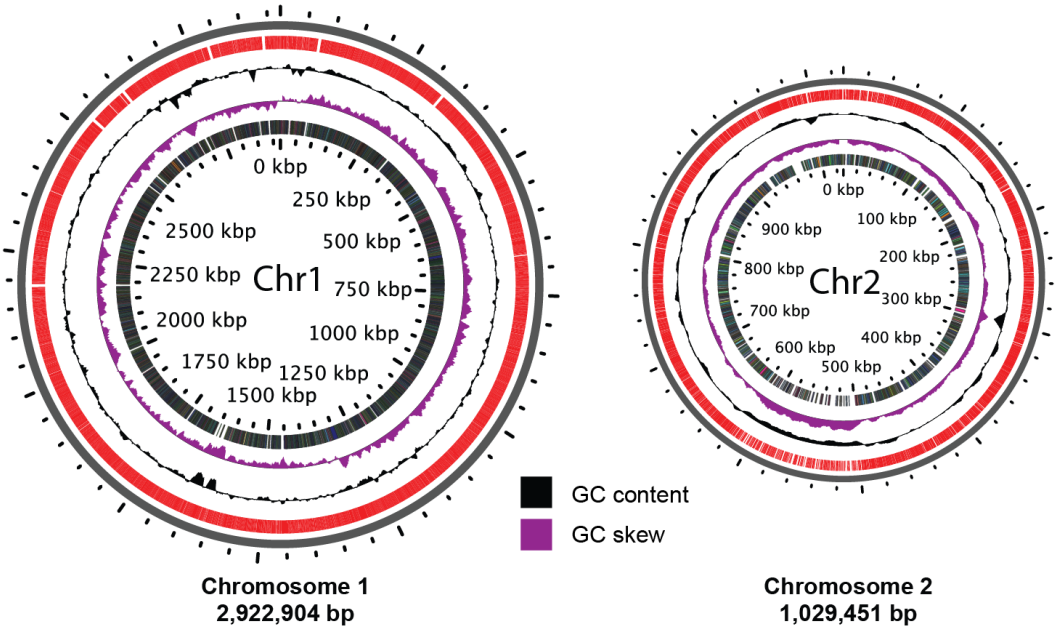

Figure 4.13 – Illustration of the NCTC 30 chromosome sequences. Modified from [424].

A comparison between the NCTC 30 assembly and that of the N16961 reference strain revealed a large inversion in NCTC 30 chromosome 1 of ~ 1,040,746 bases, between genes *VC_1056* and *VC_2013* (Figure 4.14). The short-reads from the same NCTC 30 gDNA preparation used for the assembled long-read sequencing were mapped back to both the PacBio assembly and to N16961, and to guard against this inversion being an artefact of genome assembly using long-reads only, paired reads were identified that mapped to either side of the inversion junction, and individual reads were identified which spanned the junction itself (Figure 4.15). This was confirmed further using sequencing data from a second gDNA isolation from MJD382, as well as from MJD439, a culture grown from a colony of NCTC 30 that had been separated from MJD382 at passage 1 (Methods, section 2.2.4; data not shown).

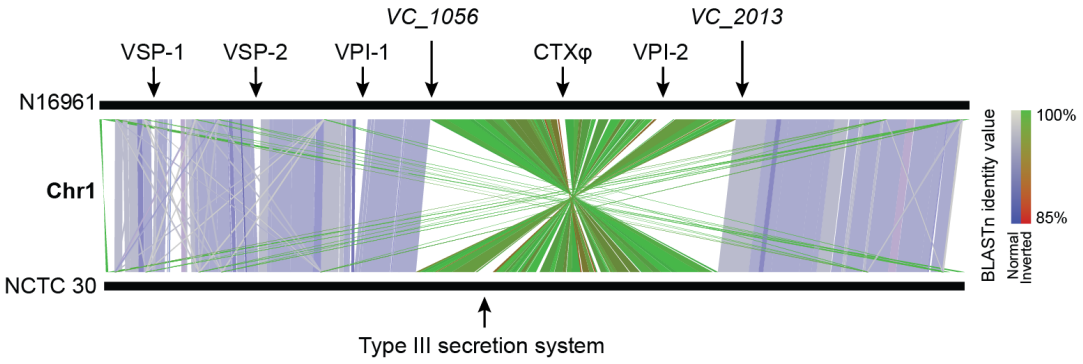

Figure 4.14 – A comparison between chromosome 1 of NCTC 30 and N16961. The chromosomal location of genomic islands and virulence determinants, discussed in subsequent sections, are indicated. The position of the inversion is indicted by the green/red inverted regions of synteny, in an 'hourglass' shape. Modified from [424].

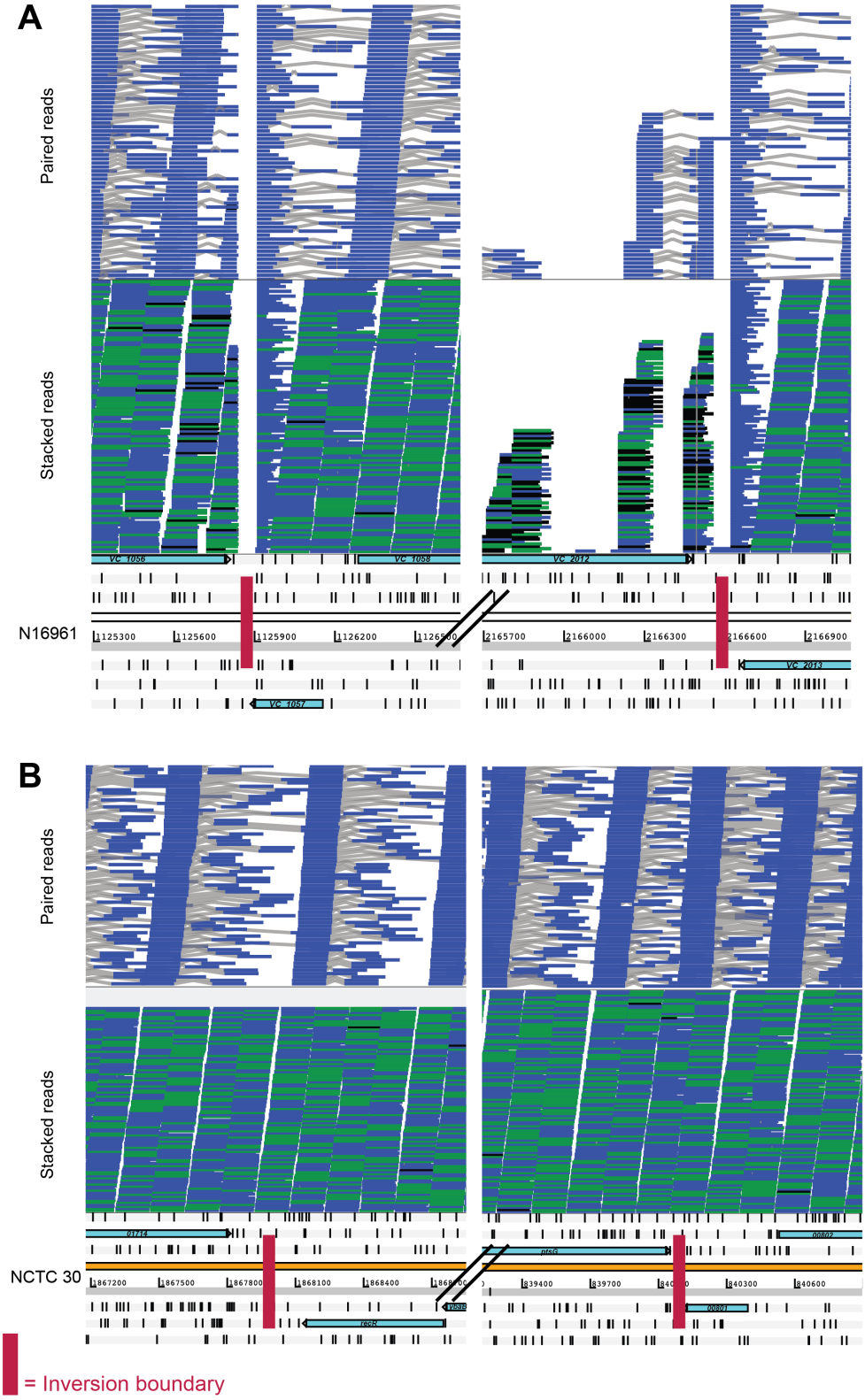

Figure 4.15 – Mapping reads across the inversion junctions. NCTC 30 short-reads were mapped to the N16961 (A) and NCTC 30 (B) genome sequences. The regions delineating the genomic inversion shown in Figure 4.14 are presented. In N16961, no reads map to the inversion region adjacent to *VC_1057*, and none of the reads that map to one of adjacent sequences have a paired read on the other side of the inversion site (A). The inversion site at *VC_2013* is similar. However, when NCTC 30 reads are mapped to the NCTC 30 long-read assembly, reads

map which span the inversion junction, and reads map to each side of the inversion which are paired to reads on the other side of the inversion site (B). The gene *VC_2012* is poorly conserved between NCTC 30 and N16961; hence, few NCTC 30 reads map to this gene. Hatch marks denote a truncation of the genome sequence view.

4.3.8 – NCTC 30 motility and flagellation defects

When NCTC 30 was being revived and cultured, it grew noticeably more slowly than other *V. cholerae* on both solid and liquid media, where it showed erratic growth kinetics (Figure 4.16). The inversion on chromosome 1 was confirmed not to encompass the *crtS* locus, and therefore should not interfere with the rate and timing of chromosome 2 replication, or with the growth rate of NCTC 30 [58] (section 1.2.1).

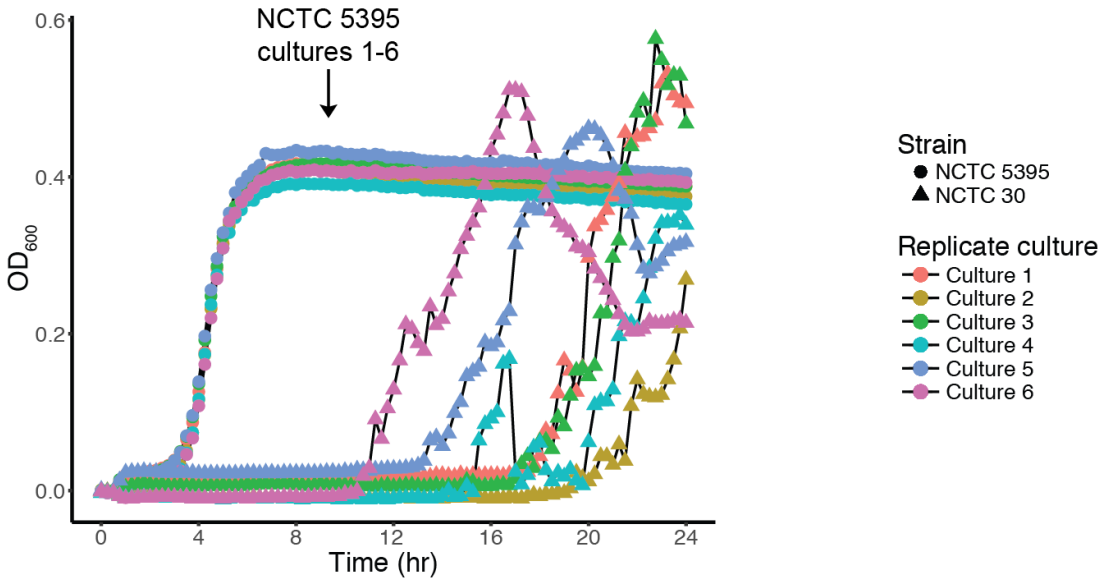

Figure 4.16 – Growth kinetics of NCTC 30 and NCTC 5395. In this assay, NCTC 30 demonstrated a growth defect at 37 ˚C relative to NCTC 5395. Under these conditions, *V. cholerae* does not grow to an OD_{600} exceeding 1.0 – accordingly, a non-logarithmic Y-axis scale has been deliberately used. Representative data from single biological experiments are reported.

This observation prompted the microscopic examination of NCTC 30, using transmission electron microscopy (TEM). Representative TEM images for NCTC 30 and a control strain of classical biotype *V. cholerae*, NCTC 10732, were captured and are presented in Figure 4.17.

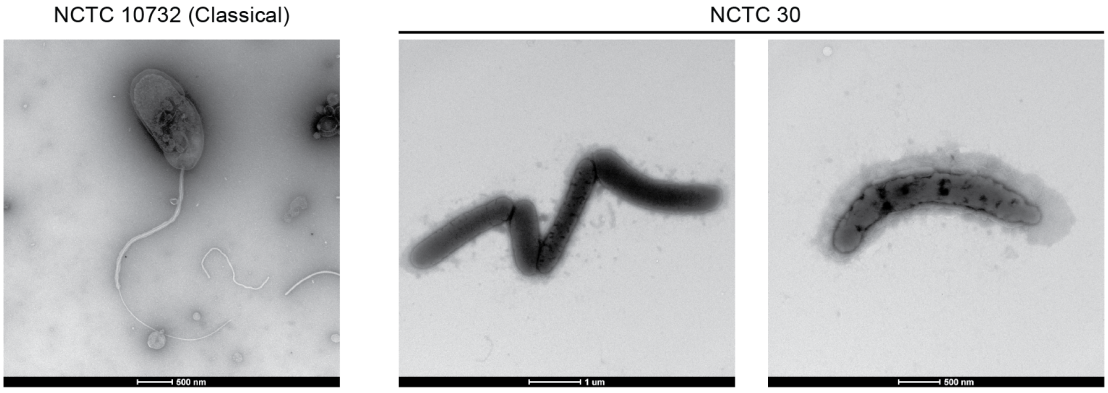

Figure 4.17 – *V. cholerae* transmission electron micrographs. NCTC 30 is compared to a control strain of flagellated *V. cholerae*, NCTC 10732. Modified from [424].

It was noted immediately that NCTC 30 cells did not appear to express the monotrichous flagellum characteristic of *V. cholerae* [1] (Figure 4.17). This was particularly intriguing because previous studies of NCTC 30 had recorded this bacterium as being flagellated [39]. Based on these microscopy data, it was hypothesised that NCTC 30 would be non-motile. Accordingly, NCTC 30 and two control strains of *V. cholerae* were cultured on motility media. No motility was observed in NCTC 30 under these conditions (Figure 4.18).

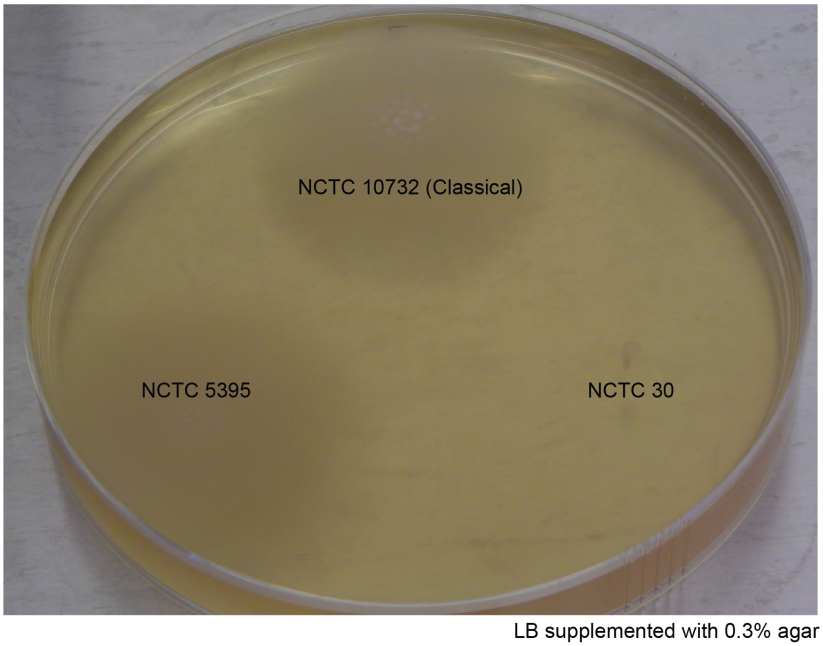

Figure 4.18 – NCTC 30 is non-motile when grown in soft agar. A lateral zone of growth corresponds to motile bacteria 'swimming' within and across the agar; such lateral growth was not observed in NCTC 30. Control strains NCTC 5395 and NCTC 10732 were used for comparisons elsewhere in this chapter. Soft agar plates were incubated for 18 hours before imaging. Reproduced from [424].

Since the flagellar regulatory and biosynthesis hierarchy has been well-studied in *V. cholerae* [426–428] (Figure 4.19), it was decided to attempt to explain this phenotype using the NCTC 30 genome sequence. Each of the genes listed in Figure 4.19 were inspected manually in the NCTC 30 genome assembly, and a putative frameshift was identified in the *flrC* gene (Figure 4.20).

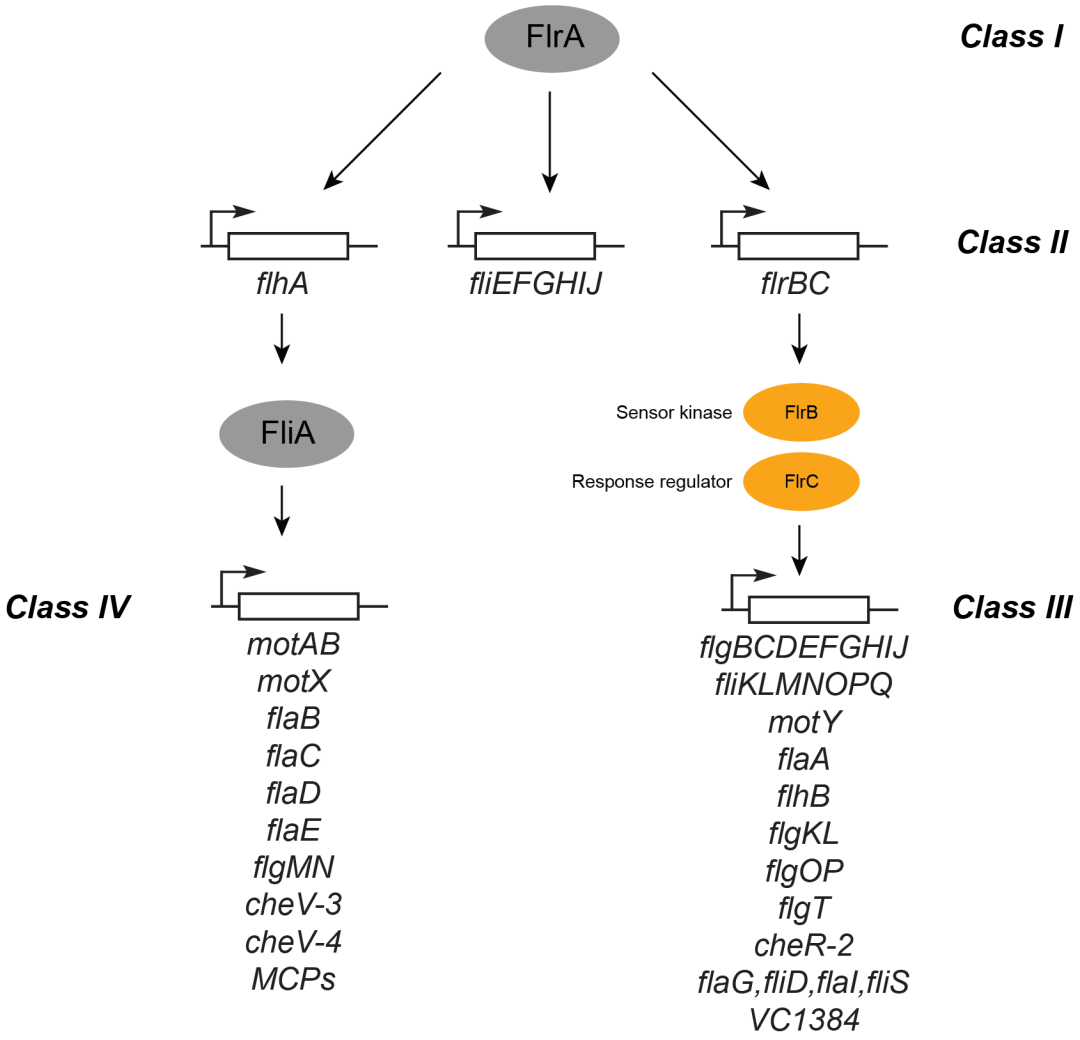

Figure 4.19 – Illustration of flagellar biosynthesis and regulatory hierarchy for *V. cholerae*. The FlrA protein activates transcription of Class II genes, including the *flrBC* operon. FlrB and FlrC then activate multiple Class III flagellar genes. Figure derived from [428].

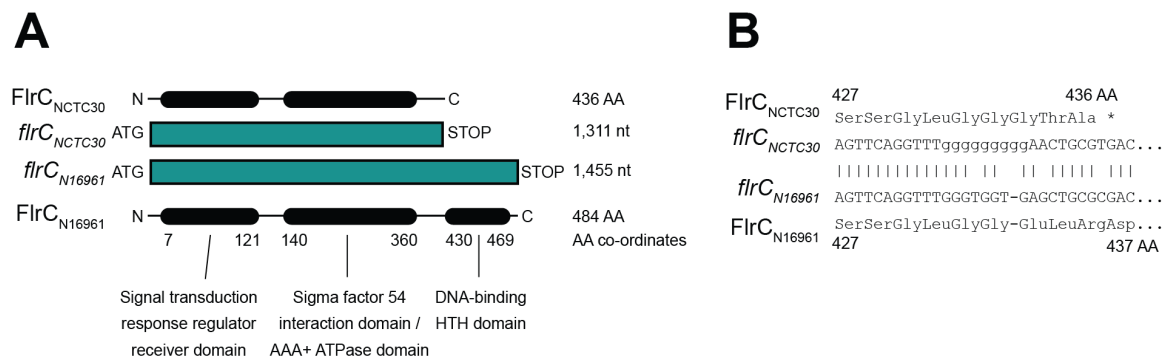

Figure 4.20 – Schematic of the predicted truncation of FlrC caused by the frameshift in *flrC*. (A): The frameshift in *flrC* is predicted to truncate FlrC, removing the DNA binding domain from this protein, thereby eliminating its ability to regulate its target genes. (B): This frameshift is predicted to disrupt the reading frame such that a premature STOP codon is introduced at amino acid position 436. Modified from [424].

The FlrC protein is a response regulator that governs the expression of Class III flagellar genes in *V. cholerae* [426] (Figure 4.19). Class III genes encode components of flagellin, as well as proteins that form part of motor and chemotaxis machinery in *V. cholerae* [427] (Figure 4.19). The mutation detected in *flrC* was predicted to introduce a frameshift, truncating FlrC by removing the last 48 amino acids from the protein (Figure 4.20). This region of FlrC is predicted to encode the DNA-binding domain, and its removal should prevent the transcription of FlrC-regulated Class III genes required for the production of *V. cholerae* flagella.

Although the hypothesis that the *flrC* frameshift was responsible for the absence of flagella was compelling, two additional sequencing technologies were used to ensure that this frameshift was not an artefact of long-read assembly. High-accuracy short-read Illumina data were mapped to the NCTC 30 assembly, and the region was also amplified and sequenced using commercial Sanger sequencing. All three sequencing technologies confirmed the existence of this frameshift in this gene (Figure 4.21).

Additionally, since Davis and Park's previous report was submitted for publication in April 1962 [39], prior to the preparation of batch 3 of NCTC 30 (Figure 4.11), it was necessary to consider the possibility that the *flrC* mutation may have arisen during the preparation of batch 3, during long-term storage [429], or during passage in our laboratory. Since the mutation was confirmed to be present in sequences of MJD382 and MJD439, two independently-sequenced clones separated from each other immediately upon rehydration of our vial of batch 3 NCTC

30, this suggested either that this mutation predated the introduction of the strain into our laboratory, or had arisen immediately upon rehydration of our lyophilised stock.

In order to test whether the *flrC* mutation was unique to our laboratory stocks, gDNA was prepared from batch 4 of NCTC 30 in laboratories at PHE, separate to our laboratory. Batch 4 was lyophilised in 1985 from a culture of batch 3 bacteria (Figure 4.11, 4.12). *flrC* was amplified from batch 4 gDNA by PCR, and Sanger sequencing was used to confirm that the *flrC* frameshift mutation was also present in batch 4 of NCTC 30 (Figure 4.21). This indicates strongly that the mutation arose either during or prior to the preparation of batch 3 of this lyophilised culture, and that this mutation ought to be present in NCTC 30 cultures which are purchased from NCTC in the future.

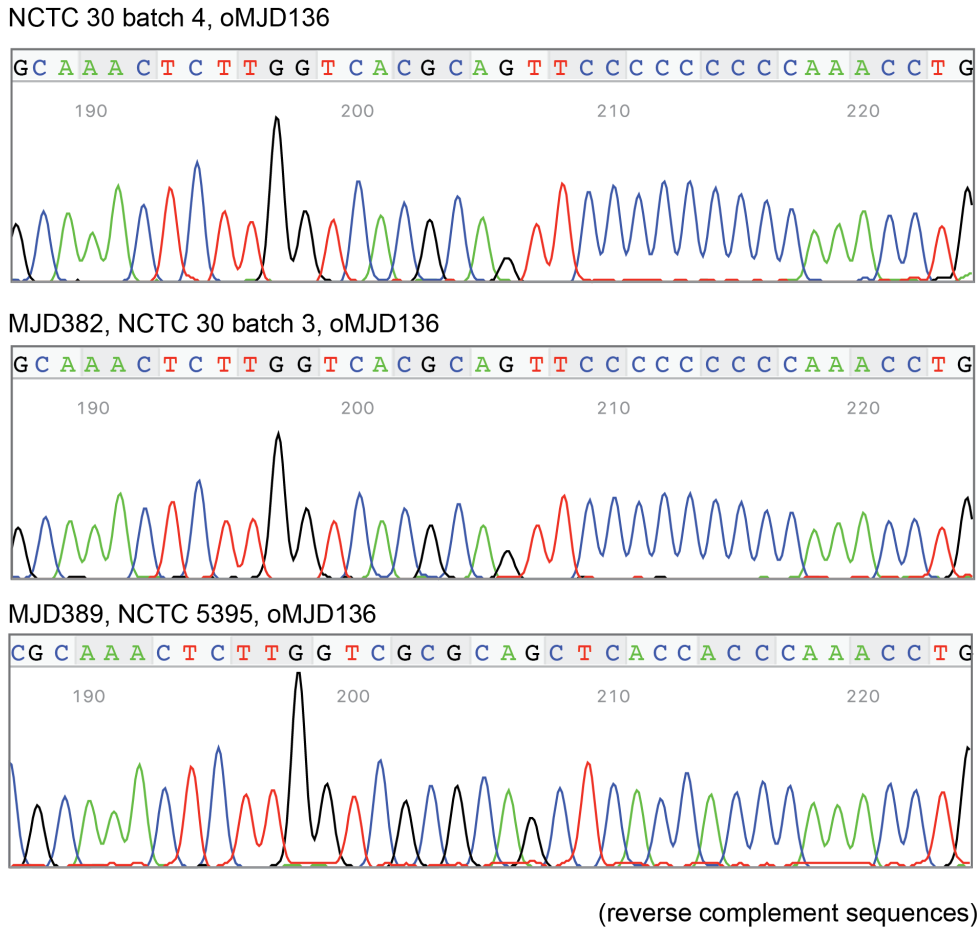

Figure 4.21 – Confirmation of *flrC* frameshift using Sanger sequencing. The mutation that causes the predicted frameshift in *flrC* in NCTC 30 batch 3 (nine G-C base pairs) are also present in the batch 4 DNA preparation. Sequencing traces were visualised using 4Peaks [14].

[14] https://nucleobytes.com/4peaks/index.html

Unfortunately, it proved impossible to culture NCTC 30 efficiently enough to attempt to transform it with plasmid vectors, either to repair the *flrC* mutation by homologous recombination, or to supply an episomal copy of *flrC in trans* to complement the defect. However, it was noted that the microscopic images of NCTC 30 were consistent with that of an *flrB* targeted mutant [426]. FlrB and FlrC form a two-component system in which FlrB phosphorylates FlrC, which then acts as a response regulator and *trans*-activates Class III genes in a σ^{54}-dependent manner [426, 430] (Figure 4.19). Given that *flrC* is the only flagellum-biosynthesis gene in NCTC 30 that is disrupted, it is reasonable to infer that this *flrC* mutation is responsible for this phenotype.

4.3.9 – Antimicrobial resistance in NCTC 30

A previous taxonomic study of *Vibrio* bacteria suggested that NCTC 30 was insensitive to a concentration of penicillin to which a control *V. cholerae* strain, NCTC 5395, was partially sensitive [39]. The NCTC 30 genome assembly was scanned for antimicrobial resistance genes *in silico* and one putative resistance gene, encoding a β-lactamase, was identified. The gene was located within the integron on the smaller chromosome, which had been fully-assembled by virtue of the long-reads. The translated sequence of this gene was used to query the NCBI nr database. The most similar sequences (99% amino acid similarity) were those of the *V. cholerae* β-lactamases CARB-7 and CARB-9. CARB-7 was first described in an environmental *V. cholerae* isolated in Argentina that resisted ampicillin to an MIC of 256 μg/ml, and was also encoded by a gene within the integron of chromosome 2 [152]. CARB-9 is also an integron-encoded β-lactamase first identified in environmental non-O1/O139 *V. cholerae* from Argentina which resisted ampicillin to an MIC of 64 μg/ml [151].

The presence of a gene predicted to encode an antimicrobial resistance determinant neither guarantees that the gene is responsible for a resistance phenotype, nor that it is actually expressed. In order to validate whether this gene was at all functional or responsible for the penicillin insensitivity reported previously [39], within the constraints of working at CL3 and with a difficult-to-culture NCTC 30 (section 4.3.8), it was decided to clone the gene into pACYC184, a low-copy vector encoding resistance to chloramphenicol and tetracycline [359]. The $bla_{CARB\text{-}like}$ gene was amplified from NCTC 30 template DNA. The amplicon was digested

and ligated into pACYC184, disrupting the native *tet* gene on the plasmid (Figure 4.22; Methods, section 2.2.12).

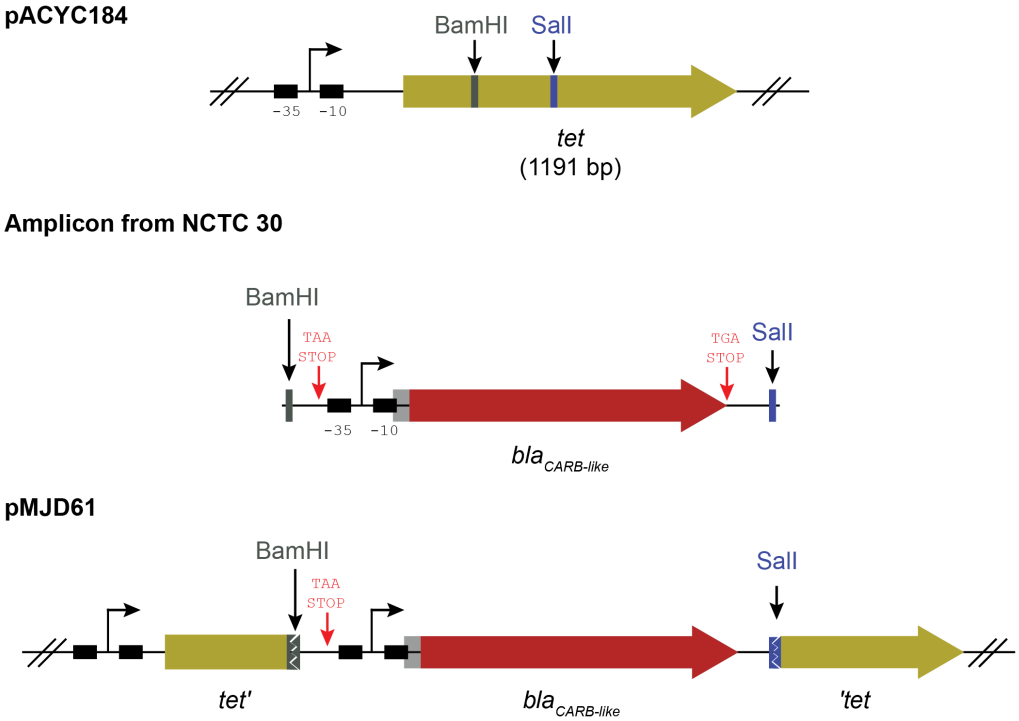

Figure 4.22 – Strategy to clone *bla*$_{CARB\text{-}like}$ from NCTC 30 gDNA. *bla*$_{CARB\text{-}like}$ was amplified from the NCTC 30 genome using primers oMJD96 and oMJD97, incorporating BamHI and SalI restriction sites. This was digested and ligated into the *tet* gene on pACYC184, introducing a premature in-frame STOP codon into *tet*. BPROM [15] predicted *E. coli* σ^{70} -35 and -10 elements within the insert (indicated). Although this software predicts promoters from *E. coli*, not *V. cholerae*, it should be noted that this might provide a native promoter from which *bla*$_{CARB\text{-}like}$ expression may be driven in *E. coli*. Figures are not to scale. Reproduced from [424].

The plasmid harbouring *bla*$_{CARB\text{-}like}$, pMJD61, was transformed into a cloning strain of *E. coli* and MICEvaluator strips were used to compare the relative sensitivities of *V. cholerae* and transformed *E. coli* (see section 2.2.13 for assay details). This assay was not used to determine quantitative MICs. Cells harbouring pUC19 which confers resistance to β-lactams, and pACYC184, an empty vector, were used as positive and negative controls respectively. NCTC 30 was found to be sensitive to ampicillin to a lesser extent than NCTC 5395, the same strain used in Davis and Park's original work [39] (Figure 4.23). The faint growth of NCTC 30 close to the test strip above the 16 µg/ml position resembles satellite colonies that emerge due to β-lactam degradation by enzyme secreted by adjacent bacterial culture (Figure 4.23).

[15] http://softberry.com

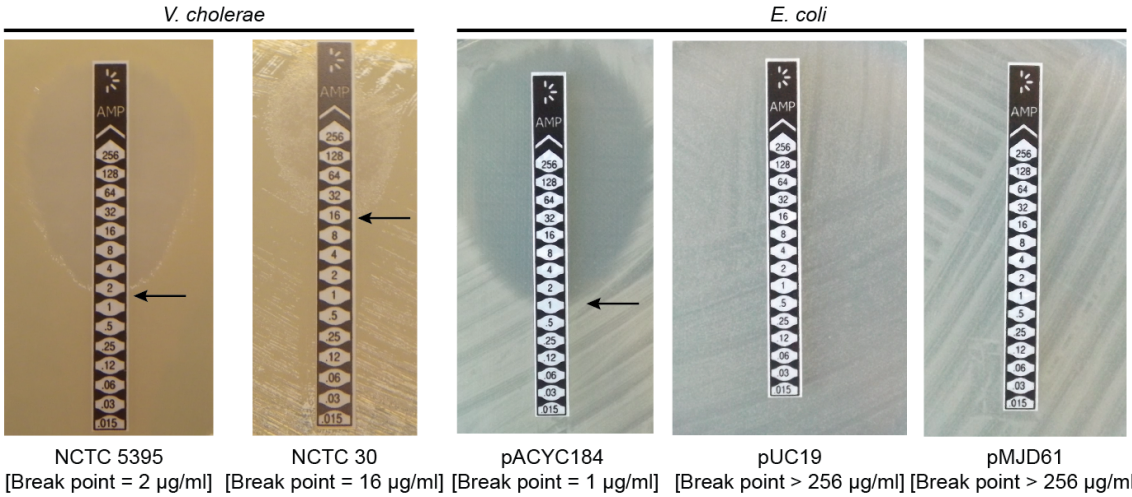

Figure 4.23 – Ampicillin sensitivity phenotypes of *V. cholerae* and plasmid-harbouring *E. coli*. Arrows indicate the break points as interpreted using the MICEvaluator manufacturer's instructions. pMJD61, the plasmid containing $bla_{CARB\text{-}like}$ (Figure 4.22), confers an equivalent ampicillin resistance to that conferred by the pUC19 ampicillin-resistance plasmid in *E. coli* 5-alpha. Modified from [424].

4.3.10 – Virulence determinants in NCTC 30

Earlier comparisons (Figure 4.14) had demonstrated the absence of the canonical pathogenicity islands VPI-1, VPI-2, VSP-1 and VSP-2, as well as the CTXφ prophage from the NCTC 30 genome. Therefore, the sequence was interrogated to attempt to identify genetic determinants that may explain the clinical symptoms giving rise to its isolation (section 4.3.7). The presence of accessory virulence genes in the assembly, including *zot*, *ace*, *hlyA*, *rtxA*, *rtxC*, *hapA*, MSHA and heat-stable enterotoxin, was determined (Table 4.4).

Accessory virulence gene	Locus ID in N16961 reference genome	Present in NCTC 30 (Percentage identity of translated protein)
Zona occludens toxin (Zot)	*VC_1458*	No (CTXφ)
Accessory cholera enterotoxin (Ace)	*VC_1459*	No (CTXφ)
Haemolysin (*hlyA*)	*VC_A0219*	Yes (98%)
Mannose-sensitive haemagglutinin (MSHA)	*VC_0398..VC_0414*	Yes, see Figure 4.24D
MARTX toxin (*rtxA*)	*VC_1451*	Yes (93%)
MARTX toxin accessory gene (*rtxC*)	*VC_1450*	Yes (100%)
HA/protease (*hapA*)	*VC_A0865*	Yes (98%)
Integrative conjugative element SXT/R391	Absent (if present, integrates into *VC_0659*)	No, *VC_0659* homologue is intact
Heat-stable enterotoxin NAG-ST (Genbank accession M85198.1)	Absent	No

Table 4.4 – Accessory virulence genes present in NCTC 30. Identity percentages were calculated by alignment of protein sequences from NCTC 30 and N16961 using BLASTp. The translated NAG-ST nucleotide sequence (accession M85198.1) was used as a tBLASTx query to scan NCTC 30 for the gene encoding this enterotoxin. Modified from [424].

Although it appeared that NCTC 30 was devoid of canonical *V. cholerae* virulence genes (Table 4.4; Figures 4.14; 4.24), it remained the case that this bacterium was of clinical origin – it had been isolated from a convalescent soldier (Figure 4.11), and it had been reported that the source of NCTC 30 was 'choleraic diarrhoea' [41]. Accordingly, the NCTC 30 genome was scanned for other putative virulence determinants seen in other *Vibrio*.

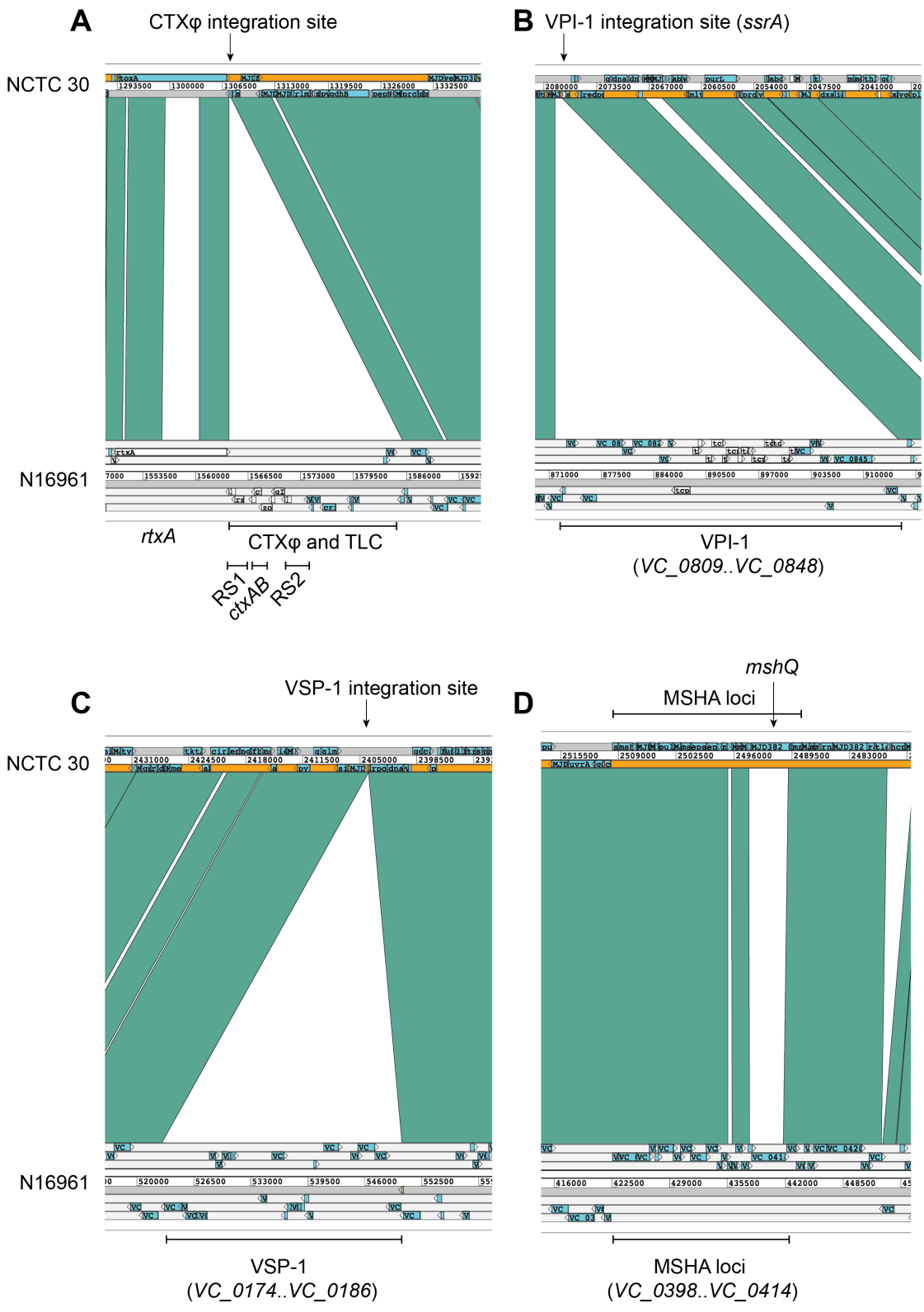

Figure 4.24 – Presence and absence of genomic islands and virulence genes in the NCTC 30 genome assembly. Comparisons are relative to the N16961 reference sequence and compare nucleotide identity. CTXφ (A), VPI-1 (B) and VSP-1 (C) are absent from NCTC 30. The genes that encode the MSHA accessory virulence determinant are present in NCTC 30 (D), although the NCTC 30 *mshQ* gene is dissimilar to that of N16961 (Table 4.4). Reproduced from [424].

A type III secretion system (T3SS) was detected on the larger chromosome of NCTC 30 (Figures 4.14; 4.25). This island is integrated between *VC_1757* and *VC_1810*, the same integration site as used by VPI-2 in N16961 (Figure 4.14, 4.25). This T3SS is more similar to the T3SS found in the genome of *V. parahaemolyticus* strain 10329 [52] than the T3SS found in *V. cholerae* AM_19226, the strain used to characterise T3SS activity in *V. cholerae* [45,46] (Figure 4.25). It is notable that a handwritten note on the NCTC's internal quality check card for NCTC 30 refers to "intermediate *V. cholerae* / *V. parahaemolyticus*" (Figure 4.11). No further information is available to explain why this note was made.

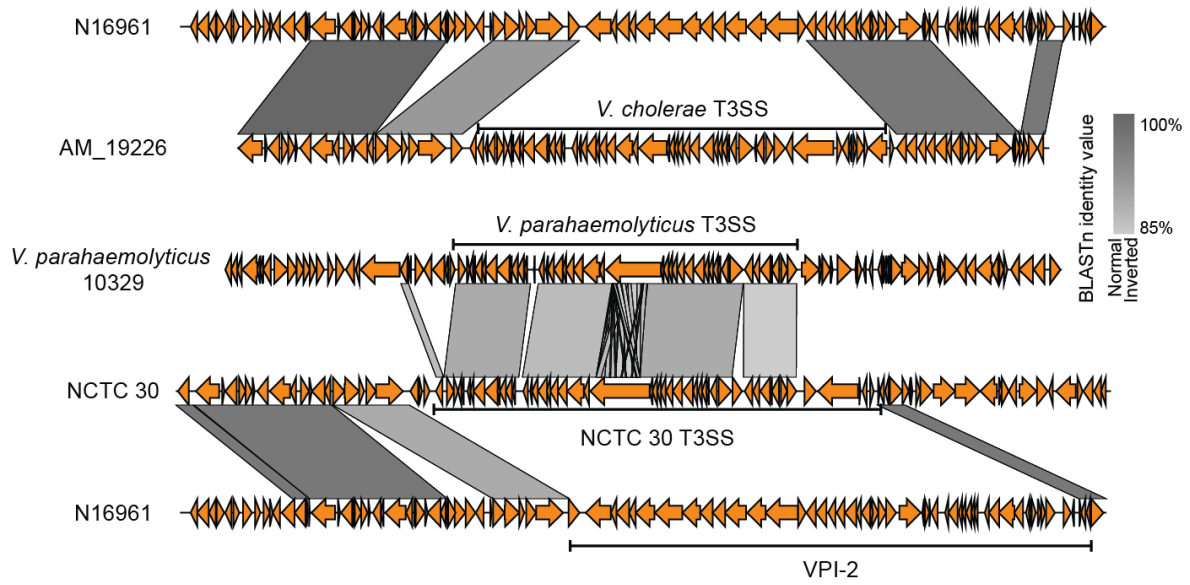

Figure 4.25 – Comparison of T3SS from NCTC 30 and other Vibrios. The T3SS harboured by NCTC 30 is most similar to one found in *V. parahaemolyticus* strain 10329, and is dissimilar to that encoded by *V. cholerae* AM_19226 [266]. The integration locus for T3SS in both NCTC 30 and AM_19226 is the same, and is the same as the VPI-2 integration locus in N16961. However, the genes flanking the T3SS in *V. parahaemolyticus* are not similar to those of *V. cholerae*. Modified from [424].

4.3.11 – Phylogenetic position of NCTC 30 and distribution of T3SS-2β

The T3SS detected in NCTC 30 and *V. parahaemolyticus* 10329 corresponds to the T3SS-2β described in limited numbers of previously-published *V. cholerae* [397] and also found in a number of the Argentinian non-O1 *V. cholerae* described in Chapter 3 (Figure 3.21). Building on the data from Chapter 3, a pangenome was constructed using 197 other *V. cholerae* genome sequences, and those of three *Vibrio* spp. that are closely related to *V. cholerae*. These context genomes are the same as those used to contextualise Argentinian non-7PET *V. cholerae* in

Chapter 3. A maximum-likelihood phylogeny was then calculated using a core-gene alignment of 2,622 genes from this pangenome. NCTC 30 was found to be more closely related to *Vibrio cholerae* than to other members of the *Vibrio* genus, though it was part of a clade separated from many of the *V. cholerae* in this collection (Figure 4.26). This observation is logical when considered together with a taxonomic study of *V. cholerae* performed in 1970, which questioned whether NCTC 30 is a true member of the *Vibrio cholerae* species [38]. The phylogenetic separation which we observed is likely to reflect the phenotypic and molecular differences that questioned the classification of NCTC 30 [38]. Intriguingly, this cluster of isolates is the same cluster of which 48853_F01, the non-toxigenic *V. cholerae* O139 isolate, was a member (Figure 4.9), and to which several of the Argentinian non-7PET *V. cholerae* belonged (Figure 3.21).

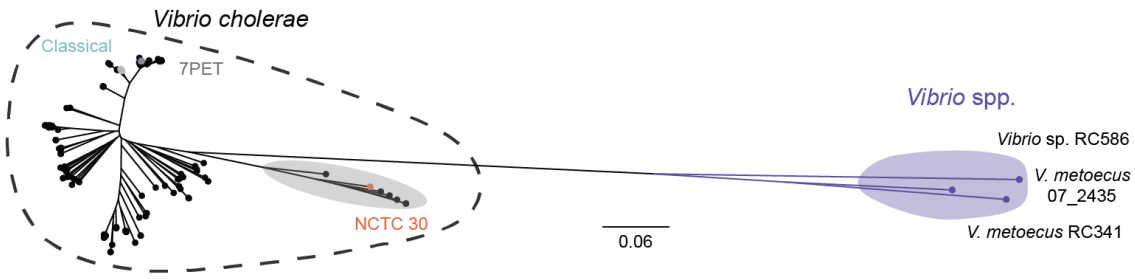

Figure 4.26 – A maximum-likelihood *V. cholerae* phylogeny including NCTC 30. An unrooted phylogeny shows that NCTC 30 clusters together with six isolates that have been previously reported to be *V. cholerae* (grey disc), including 48853_F01, the non-toxigenic *V. cholerae* O139 described in this chapter. The position of 7PET and Classical pandemic lineages [189] are noted. Scale bar denotes the number of mutations *per* variable site. Modified from [424].

Genomes in the pangenome which harboured either T3SS-2β and $bla_{CARB\text{-}like}$ were then identified (BLASTp similarity cut-offs of 95%), to map the distribution of these elements across the phylogeny. Ten genomes harbouring $bla_{CARB\text{-}like}$ homologues were identified both in strains closely related to NCTC 30 as well as in all members of the MX-3 lineage of *V. cholerae* O1, which was isolated in Mexico during 2000 [189] (Figure 4.27). Three *V. cholerae* genomes in the dataset lack CTXφ but contained genes similar to those of the NCTC 30 T3SS-2β [53]. (Figure 4.27).

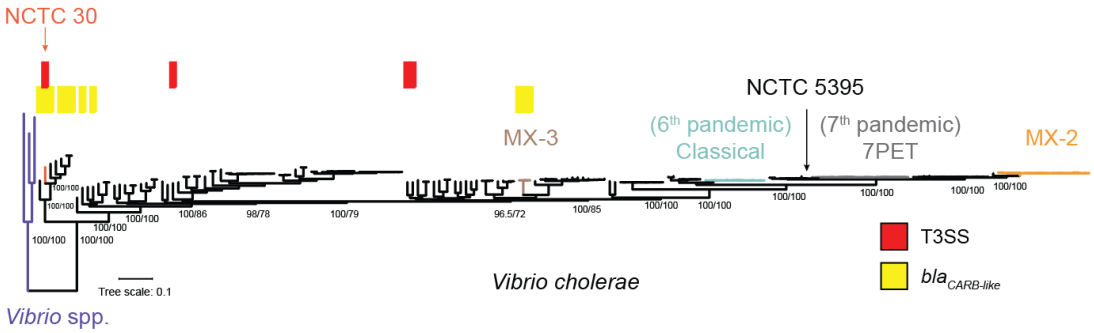

Figure 4.27 – The distribution of T3SS-2β and *bla*$_{CARB\text{-}like}$ within the *V. cholerae* phylogeny. The tree presented in Figure 4.26 is reproduced here, rooted on the three *Vibrio* spp. genomes. Lineages defined previously [189] are indicated, as is the phylogenetic position of the NCTC 5395 control strain used for molecular analyses in this chapter [213]. Modified from [424].

4.4 – Discussion

Piecing together the history of current and previous cholera pandemics requires not only an understanding of pandemic *V. cholerae* lineages, but also a view of the more diverse non-pandemic *V. cholerae* that co-exist contemporaneously with the pandemics. This theme has also been discussed throughout Chapter 3. NCTC 30 was isolated in 1916, at a time when the sixth cholera pandemic was waning [36, 163, 431], but is not a member of the Classical pandemic lineage (Figure 4.26). Very few *V. cholerae* isolates and genome sequences are available from this time period, making NCTC 30 a valuable isolate for future evolutionary studies of the *V. cholerae* species. Similarly, the non-toxigenic *V. cholerae* O139 was isolated during the seventh cholera pandemic, but was not related to the O139 sub-lineage of the 7PET pandemic lineage.

The manual and focused fine-scale analysis of single genomes presented in this chapter is highly complementary to the larger-scale analyses described in Chapter 3. The observations made from the study of these genomes can be extrapolated to larger datasets easily – for example, examining the distribution of T3SS and β-lactamases across larger collections of *V. cholerae* (Figure 4.27). Similarly, although it was by virtue of manual inspection that the coincidence of more than one *ctxB* allele in one closed *V. cholerae* O139 genome assembly was observed, once it was clear that this needed to be looked for, it was feasible to examine larger collections of *V. cholerae* O139 genomes for the coincidence of *ctxB4* and *ctxB5*, including those *V. cholerae* O139 from Domman *et al* [235]. This approach – the extrapolation of knowledge gleaned from the study of single genomes and laboratory strains into the context of larger populations – is a theme that will be re-visited in Chapter 5.

In vitro data strongly support the hypothesis that 7PET *V. cholerae* participated in homologous recombination, mediated by natural competence, to convert from serogroup O1 to O139 [144]. *V. cholerae* O139 have only caused cholera epidemics in South East Asia [244]. It is interesting to speculate that a bacterium similar to the non-toxigenic *V. cholerae* O139 sequenced in this study might have been the source of the O139 operon which was acquired by 7PET in this event. Likewise, the O139 *V. cholerae* that are distantly related to 7PET are likely to have originated from non-O1 progenitors, as suggested previously [230]. It is also evident that non-7PET *V. cholerae* O139 may harbour pathogenicity islands – among the non-7PET serogroup O139 isolates studied by Siriphap *et al* were four non-toxigenic isolates obtained in 2011, all

of which were shown to harbour VPI-2 [242]. The closed toxigenic and non-toxigenic genome sequences reported in this chapter will therefore serve as useful reference sequences for the future genomic analysis of both pandemic and non-pandemic *V. cholerae* O139.

A recurring theme from Chapters 3 and 4 is the presence of T3SS in non-O1 *V. cholerae* of clinical origin (Figure 3.21; Figure 4.27). It may be that the T3SS encoded by NCTC 30 was responsible for clinical symptoms that led to the isolation of these bacteria, not least because it was the most compelling virulence determinant identified in the genome of this bacterium. Although a detailed characterisation of this element was beyond the scope of this PhD project, this is the focus of future study. The possibility cannot also be excluded that the patient was co-infected with another pathogen in addition to NCTC 30, perhaps an O1 *V. cholerae* or another bacterium such as enterotoxigenic *E. coli* [54,55], which might also have caused "choleraic diarrhoea". Although co-infection may explain the isolation of a non-toxigenic *V. cholerae* O139 from a patient suffering from diarrhoea, it is known that even the non-toxigenic vaccine strains of *V. cholerae* O1 can elicit a diarrhoeal response in vaccine recipients [432]. Thus, the contribution of co-infections and non-toxigenic *V. cholerae* to enteric disease should be the focus of further research.

The study of $bla_{CARB\text{-}like}$ in NCTC 30 was valuable because it provided confirmation that this gene can confer ampicillin resistance, and provides genomic explanations for previous phenotypic studies of this isolate [39]. This confirmation was necessary because NCTC 30 predates the introduction of penicillin as an antibiotic, the antimicrobial activity of which was first reported by Fleming in 1929 [433]. Consequently, NCTC 30 is unlikely to have acquired its drug resistance phenotype in response to selective pressures imposed by the therapeutic use of antibiotics. It is reasonable to speculate that NCTC 30 may possess $bla_{CARB\text{-}like}$ in order to resist antibiotics found in its environment – *i.e.*, to defend itself against antibiotic-producing micro-organisms with which it might co-exist in the environment. This might also explain why NCTC 30 appears not to resist the antibiotic completely (Figure 4.23); it may be that $bla_{CARB\text{-}like}$ is expressed at levels sufficient to protect NCTC 30 from diffuse, low-concentration antibiotics present in an environment. It is also notable that β-lactams are not recommended for the treatment of cholera [11, 434], and that although a β-lactamase gene homologous to $bla_{CARB\text{-}like}$ ($bla_{CARB\text{-}2}$) was identified in the MX-3 lineage of *V. cholerae* O1, antimicrobial susceptibility testing (AST) performed on an isolate from this lineage did not lead to this isolate being classified as resistant to β-lactams [189]. This may reflect variety in β-lactam resistance

phenotypes in *V. cholerae*; bla_{CARB-2} might elevate β-lactam resistance in MX-3, but not to a level sufficient to classify a strain as "resistant" to an antimicrobial.

One of the most striking observations in this chapter is that both 48853_F01 and NCTC 30 occupy the same clade of the *V. cholerae* phylogeny, one which is very distinct from the section of the tree which containing pandemic clones and other *V. cholerae* O1 isolates (Figures 4.9, 4.26, 4.27). The phylogenetic separation which was observed is likely to reflect the phenotypic and molecular differences that led to the taxonomic classification of NCTC 30 being questioned [38], though these collective data do strongly indicate that isolates in this clade are indeed *V. cholerae* rather than another species *per se*. The structure and gene content of this clade will be explored in more detail in Chapter 5.

In Chapter 5, I will collate the genomes discussed in this chapter and the non-7PET Argentinian genomes from Chapter 3, together with an additional set of genome sequences from historical and contemporary diverse *V. cholerae*, to produce a collection of ~600 genome sequences. Using these data, I will explore the distribution of key virulence determinants across the diversity of *V. cholerae*, with particular focus on the phylogenetic clade to which 48853_F01 and NCTC 30 belong, on T3SS and accessory virulence determinants, and on the genes and SNVs which determine the El Tor biotype.