SCIENTIFIC REPORTS

OPEN

High quality reference genomes for toxigenic and non-toxigenic *Vibrio cholerae* serogroup O139

Received: 5 September 2018
Accepted: 13 March 2019
Published online: 10 April 2019

Matthew J. Dorman[1], Daryl Domman[1], Muhammad Ikhtear Uddin[2], Salma Sharmin[2], Mokibul Hassan Afrad[2], Yasmin Ara Begum[2], Firdausi Qadri[2] & Nicholas R. Thomson[1,3]

Toxigenic *Vibrio cholerae* of the O139 serogroup have been responsible for several large cholera epidemics in South Asia, and continue to be of clinical and historical significance today. This serogroup was initially feared to represent a new, emerging *V. cholerae* clone that would lead to an eighth cholera pandemic. However, these concerns were ultimately unfounded. The majority of clinically relevant *V. cholerae* O139 isolates are closely related to serogroup O1, biotype El Tor *V. cholerae*, and comprise a single sublineage of the seventh pandemic El Tor lineage. Although related, these *V. cholerae* serogroups differ in several fundamental ways, in terms of their O-antigen, capsulation phenotype, and the genomic islands found on their chromosomes. Here, we present four complete, high-quality genomes for *V. cholerae* O139, obtained using long-read sequencing. Three of these sequences are from toxigenic *V. cholerae*, and one is from a bacterium which, although classified serologically as *V. cholerae* O139, lacks the CTXφ bacteriophage and the ability to produce cholera toxin. We highlight fundamental genomic differences between these isolates, the *V. cholerae* O1 reference strain N16961, and the prototypical O139 strain MO10. These sequences are an important resource for the scientific community, and will improve greatly our ability to perform genomic analyses of non-O1 *V. cholerae* in the future. These genomes also offer new insights into the biology of a *V. cholerae* serogroup that, from a genomic perspective, is poorly understood.

Vibrio cholerae is the aetiological agent of cholera, an acute, life-threatening diarrhoea which has spread worldwide in seven pandemics since the nineteenth century. *V. cholerae* is typically sub-classified into serogroups on the basis of its somatic O-antigen. Despite there being over 200 serogroups of *V. cholerae*[1,2], only serogroup O1 has caused large scale epidemics historically[3]. Previous cholera pandemics have been caused by the classical biotype of *V. cholerae* O1, whereas the ongoing seventh pandemic, which began in the 1960s, is caused by the El Tor biotype of *V. cholerae* O1[3,4]. Non-O1 serogroups of *V. cholerae* do not appear to cause pandemics, though they may cause outbreaks of disease. This is exemplified by an outbreak in Sudan in 1968, caused by *V. cholerae* O37, which was subsequently found to be genetically related to pandemic *V. cholerae* O1[5–8].

In 1992, a *V. cholerae* clone of serogroup O139 caused a large cholera epidemic which spread rapidly across Bangladesh and India[9,10]. Due to the geographic location of the epidemic, this clone was given the name *Vibrio cholerae* O139 Bengal[10] (dubbed *V. cholerae* O139 hereafter). *V. cholerae* O139 caused substantial numbers of cholera cases in Southeast Asia in the early 1990s, and was anticipated to emerge as the aetiological agent of an eighth cholera pandemic[11–13]. However, rather than causing a pandemic, *V. cholerae* O139 was associated only with a low-level incidence of cholera cases after the initial 1992–93 epidemic, until a second large outbreak occurred in Bangladesh in the Spring of 2002[14]. The re-emergence of this serogroup renewed fear that an eighth pandemic of cholera was beginning, driven by *V. cholerae* O139[15]. Once again, *V. cholerae* O139 did not proceed to cause a cholera pandemic, and although it no longer appears to be causing epidemic cholera, this serogroup has continued to be isolated since 2002. Recently, non-toxigenic *V. cholerae* O139 have been isolated in Thailand[16]. Toxigenic *V. cholerae* O139 have been isolated in China as recently as 2013[17], and continue to be isolated in

[1]Wellcome Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, CB10 1SA, United Kingdom. [2]Infectious Diseases Division, International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh. [3]London School of Hygiene and Tropical Medicine, Keppel Street, London, WC1E 7HT, United Kingdom. Matthew J. Dorman, Daryl Domman and Muhammad Ikhtear Uddin contributed equally. Correspondence and requests for materials should be addressed to F.Q. (email: fqadri@icddrb.org) or N.R.T. (email: nrt@sanger.ac.uk)

Internal sequence ID (PacBio)	Sample Name	CTXφ present?	Accession (PacBio reads)	Accession (closed chromosomal assembly)	Accession (Illumina reads)	Genome Size (bp)	Average coverage of *de novo* assembly with long reads (X)	Coverage of N16961 (%)	Number of SNVs relative to N16961
48853_F01	MP_070116	No	ERR1716489	LT992490-LT992491	ERR568405	4123525	165.76	58.5	122865
48853_G01	P_0684000	Yes	ERR1716490	LT992486-LT992487	ERR568406	4092641	170.56	97	271
48853_H01	ICVB_2236_02	Yes	ERR1716491	LT992488-LT992489	ERR568407	4092645	147.29	97	270
48853_A02	SMIC_67_01	Yes	ERR1716492	LT992492-LT992493	ERR568408	4092644	165.65	97	274

Table 1. Summary of the *V. cholerae* O139 genome assemblies generated in this study. The accession numbers for both the long-read sequences and assemblies generated in this study, and the original short reads used for assembly polishing, SNV calling, and phylogenetic analyses (see Methods) are reported. The SNV counts reported do not account for the removal of recombinogenic sequences, since the non-toxigenic isolate was not included in the recombination analysis. Average coverage values taken from *de novo* HGAP assemblies.

Bangladesh. Six *V. cholerae* O139 strains were isolated in Bangladesh between 2013 and 2014[18] (four included in this analysis), and three more strains were isolated between 2015 and 2017.

Despite the low incidence of cholera caused by *V. cholerae* O139, this serogroup continues to be important to both the research and public health communities, not least because the disease caused by *V. cholerae* O139 is clinically indistinguishable from that caused by *V. cholerae* O1[10]. Accordingly, *V. cholerae* O139 continues to be the subject of surveillance in Southeast Asia and is included in cholera vaccine formulations[19–21]. Early genetic and biochemical studies demonstrated that O139 strains were closely related to O1 seventh pandemic El Tor strains, and it was suggested that *V. cholerae* O139 had arisen from an O1 El Tor ancestor[9,22–24]. This was subsequently confirmed using whole-genome sequencing, which showed that toxigenic *V. cholerae* O139 formed a discrete sub-lineage within the seventh pandemic El Tor (7PET) lineage[25,26].

Although the clinical diseases caused by *V. cholerae* O1 and O139 are indistinguishable, there are notable differences between *V. cholerae* O139 and *V. cholerae* O1 in addition to their serogroup. For instance, *V. cholerae* O139 expresses a polysaccharide capsule, which 7PET *V. cholerae* O1 isolates lack[27]. The capsule is encoded by genes not found in other 7PET *V. cholerae* O1 genomes, which are located adjacent to the locus encoding lipopolysaccharide (LPS) biosynthesis genes in *V. cholerae* O139[15,28–32]. It has also been reported that the complement of genomic islands found in the genome of MO10, a *V. cholerae* O139 strain, differs from that found in 7PET *V. cholerae* O1[25]. The MO10 genome sequence is currently used to represent *V. cholerae* O139 in comparative genomic analyses, but its genome sequence is incomplete and comprises 84 contigs (assembly accession number GCA_000152425.1).

V. cholerae O139 caused 93% of laboratory-confirmed cholera cases in China in the early 2000s[19]. Such data demonstrate why *V. cholerae* O139 has twice been feared to be responsible for an eighth cholera pandemic. Despite the clinical importance of this serogroup which continues to be isolated today, a closed reference genome for *V. cholerae* O139 has not yet been published. Here, we report the first high-quality, closed reference genome sequences for this *V. cholerae* serogroup. We have used long-read sequencing to obtain the complete genome sequences of four recent *V. cholerae* O139 isolates from Bangladesh[18]. Three of these are toxigenic members of the 7PET lineage, two of which were isolated from asymptomatic patients from within a household where there had been a confirmed cholera case. The fourth isolate was acquired from a patient suffering from diarrhoea, vomiting and dehydration (see ref.[18] for full details of the clinical history surrounding these four isolates), but is a non-toxigenic *V. cholerae* O139 variant that is not part of the 7PET lineage, and does not harbour the CTXφ bacteriophage[18]. This non-7PET genome offers an opportunity to study the genetic factors that enable non-toxigenic *V. cholerae* to cause diarrhoea in patients.

Results

Structure of CTXφ tandem arrays in *V. cholerae* O139. Using long-read sequencing read data, we generated single, contiguous genome sequences for the two chromosomes of each isolate sequenced in this study. Use of the corresponding short-read data for each isolate to correct these assemblies did not improve the assemblies. None of these isolates contained a third replicon, as has been reported elsewhere in other *V. cholerae*[33]. In order to estimate genetic distance, we mapped each sample's corresponding short-read data[18] to the 7PET reference genome N16961 and called single nucleotide variants (SNVs) between the mapped reads for these isolates and N16961. These SNV data are provided in Table 1, together with summary statistics for these closed assemblies. Since the three toxigenic samples were found to have near-identical genomes, varying in size by 4 bases at most, and differing by fewer than five SNVs between one another (SNVs were determined relative to N16961), 48853_H01 was selected as an exemplar sequence for further analyses.

The CTXφ bacteriophage integrates into the *V. cholerae* chromosome in a XerCD-dependent manner, by recombination between the CTXφ *attP* site and bacterial *attB* site, which produces hybrid *attL* and *attR* sequences[34,35]. This occurs after the replicative, circular form of the viral genome forms after infection of the cell, and this integration usually involves tandem integrations of CTXφ into the genome[36]. CTXφ replication is achieved by the production of ssDNA from chromosomal tandem arrays of CTXφ in a manner dependent on the CTXφ-encoded RstA protein, where RstA nicks the CTXφ replication origin located in the intergenic region Ig-1[36,37]. The exposed 3′ site permits synthesis of CTXφ DNA up until the second, tandem CTXφ replication origin is encountered, which is also a substrate for RstA cleavage, creating a free CTXφ genome[36,37]. We observed

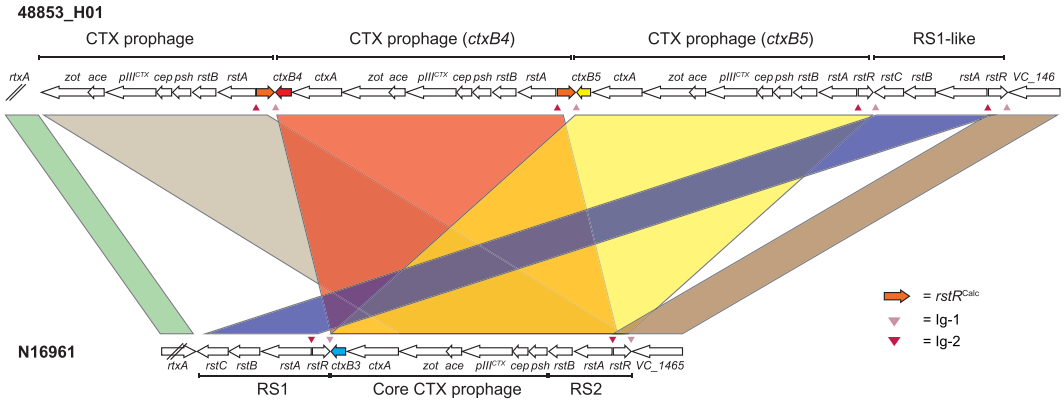

Figure 1. Tandem copies of the CTXφ bacteriophage in *V. cholerae* O139. An illustration of the genomic organisation of three tandem copies of CTXφ in the toxigenic *V. cholerae* O139 samples in this study. Two copies of the *ctxAB* operon are present in these genomes, which harbour different *ctxB* alleles to one another and to that of N16961. Exemplar data from the assembly for 48853_H01 are presented. Loci are not to scale. Figure annotation based on ref.[36].

two additional CTXφ bacteriophage sequences in tandem, relative to that found in the N16961 reference genome, located between *VC_1450* and *VC_1467* in the larger chromosome in the three toxigenic isolates (Fig. 1). A partial third repeat of CTXφ was evident, comprising the genes between and including *zot* and *rstA*, and an *rstR* open reading frame corresponding to $rstR^{Calc}$ [38] (Fig. 1). We identified a complete *attL* sequence adjacent to the *VC_1465* locus in 48853_H01[35]. The phage sequence in the *attR* site adjacent to *rtxA* is not identical to that reported by Huber and Waldor[35], although the *attR* sequence does contain both the central recombination identity sequence and the residual bacterial *attB* sequence.

Although tandem repeats of CTXφ genes in *V. cholerae* O139 have been reported previously[15,38,39], difficulty in assembling these repetitive regions with short-read sequencing data meant that these repeats were not identified in our original sequencing of these isolates[18]. We mapped the short-read data to the long-read assemblies for each of these genomes, and to the N16961 reference, to confirm that short reads mapped to both the *ctxB4* and *ctxB5* variant CTXφ regions, and that when mapped to N16961, the coverage of the CTXφ region was approximately double that of the surrounding chromosome (Supplementary Fig. S1). Manual inspection of these mapping data showed that reads from both *ctxB* alleles mapped to the N16961 *ctxB* locus.

We noted that the two *ctxB* genes in these genomes are of different alleles (Fig. 1). The *ctxB* gene closest to *rtxA* in these assemblies was a *ctxB4* allele, and the second *ctxB* was a *ctxB5* allele. Both of these are *ctxB* alleles that have been found in *V. cholerae* O139 strains previously[40,41]. The presence of more than one *ctxB* allele in the same *V. cholerae* genome has not been reported previously to our knowledge, though it has been reported that *V. cholerae* O139 can harbour more than one type of CTXφ phage simultaneously[15,38,42].

Genomic islands and antimicrobial resistance genes in *V. cholerae* O139. Having observed these unusual CTXφ configurations, we scanned the four assemblies for the presence and absence of the genomic islands that are associated with pandemic *V. cholerae*: VSP-1, VSP-2, VPI-1, VPI-2, and the drug resistance genetic element SXT[25]. We identified VPI-1 and VSP-2 islands in all three of the toxigenic *V. cholerae* O139 isolates, and we confirmed that VPI-2 is severely truncated to the point of absence, as described previously for MO10[25,43] (Fig. 2). We identified a genomic island integrated into the *VC_0659* locus (encoding peptide chain release factor 3) in each of the three toxigenic *V. cholerae* O139 assemblies, identical to SXT, which is also known as ICE*Vch*Ind4[44]. SXT is integrated into the same locus as it is in MO10. We also identified an insertion into *VC_0659* in the non-toxigenic genome assembly, which was 64% identical to ICE*Vch*Ind4 (Table 2). All of these observations agree with data from Chun *et al.*[25] on the distribution of genomic islands in the *V. cholerae* O139 MO10 genome.

In the three toxigenic *V. cholerae* O139 isolates, we detected the VSP-1 element integrated on the larger chromosome between genes *VC_0173* and *VC_0187*, as found in N16961[25] (Fig. 2). However, we also observed a sequence of DNA on the smaller chromosome of each of the toxigenic isolates, integrated between *VC_A0695* and *VC_A0696*, that was 99% identical to VSP-1 (*VC_0175* to *VC_0186*; Supplementary Fig. S2). This suggested that a second copy of the VSP-1 element was present on the second chromosome in each of these genomes. We mapped the previously-published Illumina reads for these genomes[18] to the N16961 reference genome and plotted the read depth for VSP-1 relative to the surrounding genome (Supplementary Fig. S2), which further supported the conclusion that these genomes contain a second copy of VSP-1.

This VSP-1 duplication was detected in each of the three toxigenic *V. cholerae* O139 genome assemblies (summarised in Table 2). It is known that VSP-1 is capable of excising from the larger *V. cholerae* chromosome[45], and it has been previously reported that the Matlab variant *V. cholerae* O1 strain MJ-1236 harbours a second copy of VSP-1 integrated between *VC_A0695* and *VC_A0696*[46]. Grim *et al.*[46] used PCR to identify a single clinical isolate of *V. cholerae* O139 from Bangladesh that harboured an insertion between *VC_A0695* and *VC_A0696* resembling VSP-1, but this isolate was not described further. This phenomenon is likely to be that which we have now confirmed to be present in these three *V. cholerae* O139 genomes.

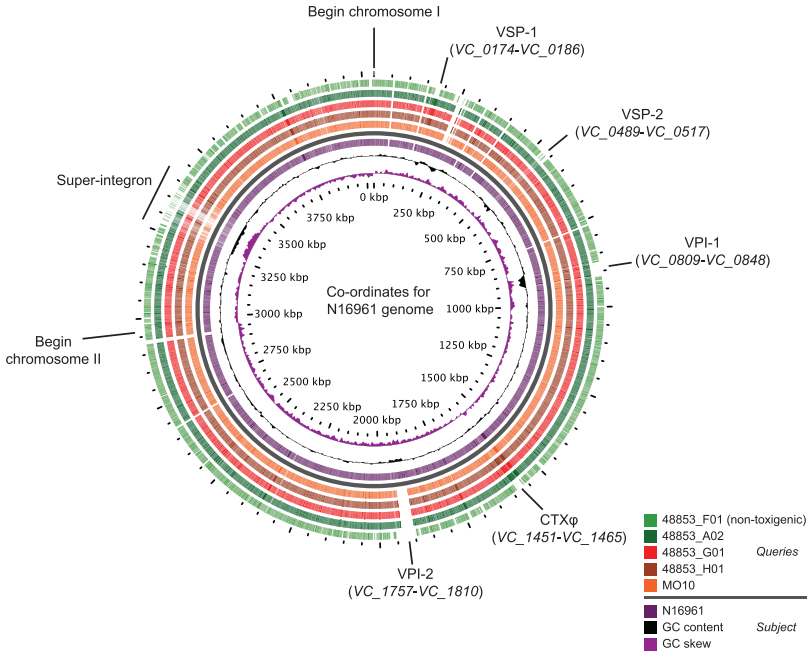

Figure 2. BLAST atlas comparing the genomes used in this study and MO10 to the N16961 reference genome. The presence of VPI-1, VSP-1, and VSP-2 in the three toxigenic *V. cholerae* O139 assemblies, as well as the truncation of VPI-2, is indicated. The non-toxigenic strain did not harbour VSP-1, VSP-2, or VPI-2 (indicated), although a region homologous to part of VPI-1 (*VC_0809* to *VC_0816*) was detected on the larger chromosome. The sequences of both *V. cholerae* chromosomes were concatenated to generate this figure; the boundary between the chromosomes is denoted.

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 (*VC_0809-VC_0816*; deletion of *VC_0817-VC_0848*)	Absent	Absent	64% match to ICE*Vch*Ind4
48853_G01	Present, and duplication of *VC_0175–0186* 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 of *VC_0175–0186* 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 of *VC_0175–0186* on chr2	Present	Present	Deletion of *VC_1761–1787*	Present, in more than one copy	100% match to ICE*Vch*Ind4

Table 2. Presence and absence of selected 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.

We also compared the genomic island complement of these four sequences with that of MO10. MO10 harbours a kappa prophage (GI-11[25]), which is absent from N16961 and also absent from the four sequences in this study (Supplementary Fig. S3). Likewise, MO10 harbours the *Vibrio* VSK prophage (GI-16[25]), which is absent from both N16961 and the O139 sequences in this study (Supplementary Fig. S3). MO10 does not appear to harbour the second VSP-1 copy on chromosome 2 which we identified in the three toxigenic isolates. The SXT variant harboured by MO10 is expanded relative to that found in these strains (Supplementary Fig. S3), and this expansion includes genes conferring resistance to the antimicrobials streptomycin (*strAB*), sulfamethoxazole (*sul2*), trimethoprim (*dfr18*), and chloramphenicol (*floR*). We scanned the assemblies for the four O139 genomes for antimicrobial resistance genes. The non-toxigenic 48853_F01 genome does not harbour any known antimicrobial resistance genes. The three toxigenic O139 genomes also do not contain any antimicrobial resistance genes, though they do harbour a *catB9* gene that is known not to confer antibiotic resistance[47]. These data are concordant with the original antimicrobial sensitivity testing of these isolates, which found that they were resistant only to nalidixic acid[18]. We confirmed that these four isolates harbour an S83I mutation in GyrA, and that 48853_F01 also contains an A171S mutation in GyrA and a S85L mutation in ParC. All of these mutations are associated with nalidixic acid resistance in *V. cholerae*[48]. We also scanned the assembled genomes of the four isolates for the presence of *V. cholerae* accessory virulence genes, to determine whether candidate virulence genes were present in the genome of the otherwise non-toxigenic *V. cholerae* O139 isolate[18] (Table 3). We did not identify any virulence determinants in the 48853_F01 genome assembly other than those typically found in *V. cholerae*[3].

Accessory virulence gene (N16961 locus ID or accession number)	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 3. Presence and absence of accessory virulence genes in *V. cholerae* O139 genome assemblies. Gene presence and absence was determined using ACT[67] to visualise BLASTn synteny plots, and using tBLASTx to scan assemblies using the NAG-ST nucleotide sequence as a query.

Phylogenetic analysis. We constructed a maximum-likelihood phylogeny from an alignment of core genes from 65 diverse *V. cholerae* genomes, and confirmed that despite its serogroup, 48853_F01 is not a member of the 7PET O139 sublineage (Fig. 3A). We did find that the three toxigenic genomes were members of 7PET, and we used the previously-published short-read data for these genomes to place these isolates into phylogenetic context with 114 other *V. cholerae*, including 23 O139 genome sequences[16,18,26] (117 genomes in total; Supplementary Table S1). We found that the three toxigenic isolates in this study clustered together with other 7PET *V. cholerae* O139 sequences from Bangladesh and India from 1992 to 2002 (Fig. 3B). The closest relatives of these three strains, which were isolated in 2013 and 2014, were isolated from Bangladesh in 2002 (A383, Case_09–12). All of these were also closely related to *V. cholerae* O139 samples from 1992–1995, including a recently-sequenced collection of *V. cholerae* O139 from Thailand[16]. These results recapitulate and reinforce previously-published data[18], adding to the utility of these genomes as reference sequences for toxigenic *V. cholerae* O139. Moreover, there are no complete genome assemblies for non-toxigenic *V. cholerae* O139. Given that this isolate is clearly distinct from the 7PET O139 sublineage, we anticipate that this genome sequence will enable comparative genomic studies of *V. cholerae* other than 7PET isolates.

The previous report of these genome sequences, obtained using short-read technology alone, examined the structure of the capsule and lipopolysaccharide (LPS) synthesis loci. These analyses were performed using incompletely-assembled genome sequences[18]. We used the closed sequences obtained in this project to compare these loci across the strains in this study to N16961 and MO10 (Supplementary Fig. S4). We confirmed that the three toxigenic strains contain O139 LPS operons that strongly resemble that found in MO10 (Supplementary Fig. S4). 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[18]. In our phylogenetic analyses, we noted that 48853_F01 clustered with two Haitian non-O1 *V. cholerae* isolates from 2010 and a Mexican isolate from 1991 for which there are no serotype data (Fig. 3A; Supplementary Table S1). We found that, although these three non-O1 *V. cholerae* share capsule biosynthesis genes with 48853_F01, they do not harbour the same LPS operon (Supplementary Fig. S4). These Haitian and Mexican isolates therefore are unlikely to be *V. cholerae* of serogroup O139.

Discussion

It is essential to have complete and accurate reference sequences to perform bacterial genomic analysis. Although several studies have provided closed *V. cholerae* sequences[49–53], none to date have provided reference sequences for *V. cholerae* O139. The sequences in this study will serve as an important community resource in future studies of *V. cholerae* genomics and phylogenetics. For example, access to the closed sequences of the O139 LPS and capsule biosynthesis operons from these four strains means that it should be possible to serotype *V. cholerae* O139 sequences *in silico*. The fact that the LPS biosynthesis loci in these toxigenic and non-toxigenic strains are similar but not identical (Supplementary Fig. S4), and that the non-toxigenic strain is distantly related to the toxigenic *V. cholerae* O139 in this study (Fig. 3), suggests that there may be more than one genetic configuration that confers an O139 serogroup phenotype. In the absence of candidate virulence genes, putative or otherwise, we also cannot exclude the possibility that the non-toxigenic isolate was obtained from a patient who was co-infected with another toxigenic organism such as enterotoxigenic *Escherichia coli*[54–56].

Many of the observations in this study could only be made because of the resolution offered by long-read sequencing. For example, the observation that several *ctxB* alleles can co-exist in a single genome is striking. The co-existence of several CTXφ sequences in tandem has been reported before, such as in the O395 classical reference sequence[52,57] and in the PA1849 second-pandemic classical isolate[4]. However, in PA1849, the tandem bacteriophages are of the same *ctxB* allele, and in O395, the CTXφ array on the larger chromosome consists of one intact CTXφ and one partial prophage sequence[57]. Although these *V. cholerae* O139 had been sequenced previously, it had not been possible to assemble these genomes fully with the short-read Illumina technology used at the time. Consequently, in all three assemblies, CTXφ was not assembled into a single contig, and only one of the two *ctxB* genes was identifiable (our Illumina assemblies for 48853_G01 and 48853_H01 contain a *ctxB4* allele in

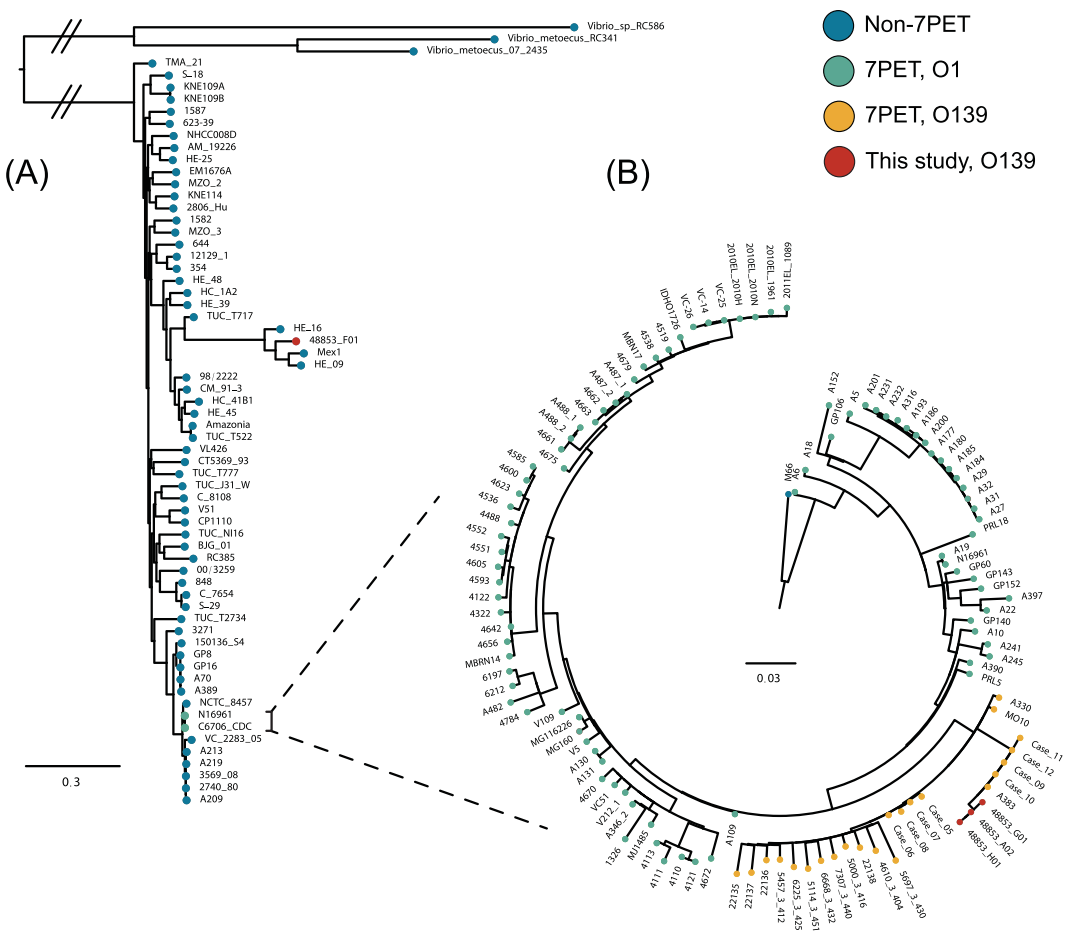

Figure 3. Phylogenetic analysis. (**A**) A total of 168,476 variable sites from a core-gene alignment of 2,103 genes from 65 genomes were used to generate a maximum-likelihood phylogenetic tree of the *V. cholerae* species, rooted on *Vibrio metoecus* and *Vibrio sp*. RC586. Two genomes that are representative of the 7PET lineage were included. The non-toxigenic isolate 48853_F01 clustered together with other non-toxigenic, non-O1/O139 *V. cholerae*. (**B**) A maximum-likelihood phylogeny of the 7PET lineage constructed using 1,629 non-recombinant variable sites across 117 *V. cholerae* genomes, rooted on M66-2. The three toxigenic samples in this study clustered together with other toxigenic O139 genome sequences, all of which form a discrete sub-lineage within 7PET. Hatch marks denote branches that were shortened artificially for illustrative purposes; an unedited tree is presented in Supplementary Fig. S5. Trees were also computed using an ascertainment bias correction model (see Methods). These are presented in Supplementary Fig. S6.

a small contig, and 48853_A02 contains *ctxB5* in a larger contig). In future studies of *V. cholerae* O139, mapping sequencing reads against these reference sequences will address this problem, which will not be resolved if data are exclusively mapped to N16961 or related sequences.

Furthermore, we note that the oligonucleotides used for PCR based *ctxB* typing[40] are 100% homologous to regions upstream and downstream of both *ctxB* loci in these three genomes. It would therefore not be possible to discriminate between *ctxB* types based on Sanger sequencing of these amplicons. This suggests that caution should be used in the interpretation of PCR-based *ctxB* typing data in the epidemiological study of cholera outbreaks, particularly if this CTXφ configuration is present in other *V. cholerae* lineages.

The functions of VSP-1 and VSP-2 are not fully understood, although it is well-accepted that these two genomic islands are found in *V. cholerae* which are members of the seventh pandemic[25]. It is known that DncV, encoded by VSP-1, represses *V. cholerae* chemotaxis[58] and that repression of chemotaxis has been linked to improved intestinal colonisation by *V. cholerae*[59]. DncV is also known to be upregulated under conditions of gastrointestinal infection, in response to conditions that activate the ToxT transcription factor *via* the TarB small RNA which prevents the production of VPI-1-encoded VspR. It is interesting to speculate that the duplication of VSP-1 might further attenuate *V. cholerae* chemotaxis under colonisation conditions *via* a gene dosage effect, thereby modulating the ability of these *V. cholerae* O139 strains to colonise the intestine.

Here, by sequencing *V. cholerae* O139 using long-read technology, we have highlighted genomic features that emphasise the genetic distinctions between *V. cholerae* O139 and *V. cholerae* O1. We have identified differences between these recent strains, N16961, and MO10, the *V. cholerae* O139 strain used for previous comparative analyses. We have also described unusual phenomena in *V. cholerae* O139 genome biology – namely, the co-existence

of more than one *ctxB* allele, and the cross-chromosome duplication of VSP-1. There also appears to have been genetic changes within *V. cholerae* O139 that has occurred since its first identification in 1992, typified by the MO10 isolate. Given that *V. cholerae* O139 is a member of 7PET, has several characteristics of a sublineage with the potential to cause pandemic disease, and continues to be isolated in recent years, research into this serogroup should continue. These reference sequences enable such research, and as well as providing interesting insights into the genome structure of recent *V. cholerae* O139, these sequences are an important resource for future genomic studies of *V. cholerae* as a pathogen and as a species.

Methods

Isolates and sequences used in this study. Four previously-described *V. cholerae* O139 isolates[18] were selected for re-sequencing on the PacBio RSII platform. A set of 178 genomes in addition to these were included in comparative genome analyses (182 genomes in total; Supplementary Table S1).

DNA isolation. Genomic DNA was prepared from 25 ml cultures of bacterial isolates grown overnight at 37 °C in LB media. Cells were harvested by centrifugation and resuspended in 2.0 ml of 25% w/v sucrose in TE buffer (10 mM Tris pH 8.0, 1 mM EDTA pH 8.0). Nuclei Lysis Solution (Promega, #A7941, 6.0 ml) was added and samples were lysed by incubation at 80 °C for five minutes. Samples were mixed with proteinase K (250 μg/ml final concentration), RNase A Solution (Promega, #A797A, 15 μg/ml final concentration), EDTA pH 8.0 (25 mM final concentration) and aqueous SDS solution (0.3% final concentration). Mixtures were incubated on ice for two hours and then at 50 °C overnight. Following enzymatic treatment, TE buffer was added to each sample (12 ml final volume). DNA was then isolated by phenol-chloroform extraction. Nucleic acids were precipitated in absolute ethanol, washed in ethanol (70% v/v), and resuspended in approximately 350 μl Tris (10 mM; pH 8.0). EDTA was omitted from the resuspension solution, to avoid interference with PacBio sequencing chemistry.

Long-read sequencing. SMRTbell libraries were created from approximately 10 μg DNA according to the manufacturer's protocol (15 kb library size, no size selection). Long reads were generated by sequencing on the PacBio RSII platform using polymerase version P6 and C4 sequencing chemistry. Sequence reads were assembled using HGAP v3[60] of the SMRT analysis software v2.3.0. The fold coverage to target when picking the minimum fragment length for assembly was set to 30 and the approximate genome size was set to 3 Mbp. The HGAP assembler assembled the reads from sample 48853_G01 into three contigs. However, assembling this sample with Canu v1.1[61] produced an assembly of two contigs, one per chromosome, and this was used for subsequent analysis. Assemblies were circularised using Circlator v1.1.3[62] and the pre-assembled reads (also known as corrected reads). The circularised assemblies were polished using the PacBio RS_Resequencing protocol and Quiver v1 of the SMRT analysis software v2.3.0. Automated annotation was performed using Prokka v1.11[63] and genus specific databases from RefSeq[64]. Pilon v1.19[65] did not identify any SNVs in any of the PacBio assemblies using the corresponding short-read data – accordingly, no short-read corrections were made to these assemblies. Raw sequencing reads and the genome assemblies described in this study have been deposited into the European Nucleotide Archive (Table 1; Supplementary Table S1).

Comparative genomics and BLAST atlas construction. The four annotated genome assemblies were compared to one another, and to the N16961 and MO10 reference genomes (see Supplementary Table S1 for accession numbers), using BLASTn[66]. These comparisons were visualised using ACT[67] and by BLAST atlas comparison using the GView web server (https://server.gview.ca/).

Read alignment, SNV identification, and core gene alignment. Paired-end Illumina reads from 116 7PET *V. cholerae* O1 and O139 samples, together with the M66-2 pre-pandemic strain (117 genomes in total; Supplementary Table S1), were mapped to the *V. cholerae* O1 El Tor N16961 reference genome (see Supplementary Table S1 for accession numbers) using SMALT v.0.7.4. Variable sites were identified using samtools mpileup v0.1.19, with parameters "-d 1000 –DsugBf", and bcftools v0.1.19[68]. High quality SNVs were determined as previously described[69], and putative recombinant regions were detected and filtered from the alignment using Gubbins[70], to produce a final alignment of 1,629 SNVs.

Prokka-annotated assemblies for 63 non-7PET and two 7PET genomes[63,71] were used to generate a species-level pan-genome using Roary[72] with the following arguments: "-e–mafft -s -cd 97". Poorly-aligned and gap-rich sites were removed from an alignment of 2,103 core gene sequences using trimAl v1.4.rev5[73] with the "-automated1" argument. A total of 168,476 variable sites were identified in the resultant alignment using SNP-sites v2.3.2[74].

Phylogenetic analysis. Maximum likelihood phylogenetic trees were constructed using RAxML v8.2.8[75] under the GTR model with the gamma distribution to model site heterogeneity (GTRGAMMA), using 500 bootstrap replicates. An alignment composed of 1,629 non-recombinant variable sites was used to generate a reference-based 7PET phylogeny. An alignment of 168,476 variable sites from 2,103 core genes was used to generate a core-gene *V. cholerae* species phylogeny. Trees were also computed using the GTR + ASC model in IQ-Tree v1.5.5[76], optimised for an input containing no invariant nucleotides, and were supported by 5,000 ultrafast bootstrap approximations and approximate likelihood ratio tests[77–79]. Phylogenetic trees were visualised using FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/) and the interactive tree of life (iTOL) v3[80].

Identification of antimicrobial resistance genes. Genome assemblies were scanned for the presence of antimicrobial resistance genes using the ResFinder web server v2.1 (https://cge.cbs.dtu.dk/services/ResFinder/)[81].

Data Availability

Sequencing reads generated during this project have been deposited into the European Nucleotide Archive (ENA; http://www.ebi.ac.uk/ena) under study accession number PRJEB14661. Assembled genome sequences have been deposited under accession numbers LT992486- LT992493. Sequence alignments, phylogenetic tree data, and other supporting information are available in Figshare (https://doi.org/10.6084/m9.figshare.6480266).

References

1. Shimada, T. *et al*. Extended serotyping scheme for *Vibrio cholerae*. *Curr. Microbiol.* **28**, 175–178 (1994).
2. Chapman, C. *et al*. Scanning the landscape of genome architecture of non-O1 and non-O139 *Vibrio cholerae* by whole genome mapping reveals extensive population genetic diversity. *PLoS ONE* **10**, e0120311, https://doi.org/10.1371/journal.pone.0120311 (2015).
3. Kaper, J. B., Morris, J. G. & Levine, M. M. Cholera. *Clin. Microbiol. Rev.* **8**, 48–86 (1995).
4. Devault, A. M. *et al*. Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. *N. Engl. J. Med.* **370**, 334–340 (2014).
5. O'Shea, Y. A., Reen, F. J., Quirke, A. M. & Boyd, E. F. Evolutionary genetic analysis of the emergence of epidemic *Vibrio cholerae* isolates on the basis of comparative nucleotide sequence analysis and multilocus virulence gene profiles. *J. Clin. Microbiol.* **42**, 4657–4671 (2004).
6. Zinnaka, Y. & Carpenter, C. C. An enterotoxin produced by noncholera vibrios. *Johns Hopkins Med. J.* **131**, 403–411 (1972).
7. Bik, E. M., Gouw, R. D. & Mooi, F. R. DNA fingerprinting of *Vibrio cholerae* strains with a novel insertion sequence element: a tool to identify epidemic strains. *J. Clin. Microbiol.* **34**, 1453–1461 (1996).
8. Beltrán, P. *et al*. Genetic diversity and population structure of *Vibrio cholerae*. *J. Clin. Microbiol.* **37**, 581–590 (1999).
9. Albert, M. J. *Vibrio cholerae* O139 Bengal. *J. Clin. Microbiol.* **32**, 2345–2349 (1994).
10. Cholera Working Group, icddr,b, *et al*. Large epidemic of cholera-like disease in Bangladesh caused by *Vibrio cholerae* 0139 synonym Bengal. *The Lancet* **342**, 387–390 (1993).
11. Finkelstein, R. A. Cholera, *Vibrio cholerae* O1 and O139, and other pathogenic Vibrios. *Medical Microbiology* [Baron, S. (ed.)] (University of Texas Medical Branch at Galveston 1996).
12. WHO. WHO | 1998 - Cholera - *Vibrio cholerae* O139 strain. *WHO* Available at: http://www.who.int/csr/don/1998_09_22/en/ (Accessed: 3rd January 2018)
13. Faruque, A. S., Fuchs, G. J. & Albert, M. J. Changing epidemiology of cholera due to *Vibrio cholerae* O1 and O139 Bengal in Dhaka, Bangladesh. *Epidemiol. Infect.* **116**, 275–278 (1996).
14. Faruque, S. M. *et al*. Reemergence of epidemic *Vibrio cholerae* O139, Bangladesh. *Emerg. Infect. Dis.* **9**, 1116–1122 (2003).
15. Faruque, S. M. *et al*. Emergence and evolution of *Vibrio cholerae* O139. *Proc. Natl. Acad. Sci. USA* **100**, 1304–1309 (2003).
16. Siriphap, A. *et al*. Characterization and genetic variation of *Vibrio cholerae* isolated from clinical and environmental sources in Thailand. *PLoS ONE* **12**, e0169324, https://doi.org/10.1371/journal.pone.0169324 (2017).
17. Yi, Y. *et al*. Genome sequence and comparative analysis of a *Vibrio cholerae* O139 strain E306 isolated from a cholera case in China. *Gut Pathog.* **6**, 3 (2014).
18. Chowdhury, F. *et al*. *Vibrio cholerae* serogroup O139: Isolation from cholera patients and asymptomatic household family members in Bangladesh between 2013 and 2014. *PLoS Negl. Trop. Dis.* **9**, e0004183, https://doi.org/10.1371/journal.pntd.0004183 (2015).
19. WHO. WHO | Weekly Epidemiological Record, 30 July 2004, vol. 79, 31 (pp 281–288). (2004). Available at: http://www.who.int/wer/2004/wer7931/en/. (Accessed: 5th July 2017).
20. WHO | Weekly Epidemiological Record, 25 August 2017, vol. 92, 34 (pp. 477–500). *WHO* Available at: http://www.who.int/wer/2017/wer9234/en/. (Accessed: 28th January 2018)
21. Saha, A. *et al*. Safety and immunogenicity study of a killed bivalent (O1 and O139) whole-cell oral cholera vaccine Shanchol, in Bangladeshi adults and children as young as 1 year of age. *Vaccine* **29**, 8285–8292 (2011).
22. Berche, P. *et al*. The novel epidemic strain O139 is closely related to the pandemic strain O1 of *Vibrio cholerae*. *J. Infect. Dis.* **170**, 701–704 (1994).
23. Calia, K. E., Waldor, M. K. & Calderwood, S. B. Use of representational difference analysis to identify genomic differences between pathogenic strains of *Vibrio cholerae*. *Infect. Immun.* **66**, 849–852 (1998).
24. Hall, R. H., Khambaty, F. M., Kothary, M. H., Keasler, S. P. & Tall, B. D. *Vibrio cholerae* non-O1 serogroup associated with cholera gravis genetically and physiologically resembles O1 E1 Tor cholera strains. *Infect. Immun.* **62**, 3859–3863 (1994).
25. Chun, J. *et al*. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc. Natl. Acad. Sci.* **106**, 15442–15447 (2009).
26. Mutreja, A. *et al*. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
27. Weintraub, A. *et al*. *Vibrio cholerae* O139 Bengal possesses a capsular polysaccharide which may confer increased virulence. *Microb. Pathog.* **16**, 235–241 (1994).
28. Bik, E. M., Bunschoten, A. E., Gouw, R. D. & Mooi, F. R. Genesis of the novel epidemic *Vibrio cholerae* O139 strain: evidence for horizontal transfer of genes involved in polysaccharide synthesis. *EMBO J.* **14**, 209–216 (1995).
29. Sozhamannan, S. *et al*. Cloning and sequencing of the genes downstream of the *wbf* gene cluster of *Vibrio cholerae* serogroup O139 and analysis of the junction genes in other serogroups. *Infect. Immun.* **67**, 5033–5040 (1999).
30. Stroeher, U. H., Parasivam, G., Dredge, B. K. & Manning, P. A. Novel *Vibrio cholerae* O139 genes involved in lipopolysaccharide biosynthesis. *J. Bacteriol.* **179**, 2740–2747 (1997).
31. Yamasaki, S., Garg, S., Nair, G. B. & Takeda, Y. Distribution of *Vibrio cholerae* O1 antigen biosynthesis genes among O139 and other non-O1 serogroups of *Vibrio cholerae*. *FEMS Microbiol. Lett.* **179**, 115–121 (1999).
32. Waldor, M. K., Colwell, R. & Mekalanos, J. J. The *Vibrio cholerae* O139 serogroup antigen includes an O-antigen capsule and lipopolysaccharide virulence determinants. *Proc. Natl. Acad. Sci. USA* **91**, 11388–11392 (1994).
33. Okada, K. *et al*. Characterization of 3 megabase-sized circular replicons from *Vibrio cholerae*. *Emerg. Infect. Dis. J.* **21**, 1262 (2015).
34. McLeod, S. M. & Waldor, M. K. Characterization of XerC- and XerD-dependent CTX phage integration in *Vibrio cholerae*. *Mol. Microbiol.* **54**, 935–947 (2004).
35. Huber, K. E. & Waldor, M. K. Filamentous phage integration requires the host recombinases XerC and XerD. *Nature* **417**, 656 (2002).
36. McLeod, S. M., Kimsey, H. H., Davis, B. M. & Waldor, M. K. CTXφ and *Vibrio cholerae*: exploring a newly recognized type of phage–host cell relationship. *Mol. Microbiol.* **57**, 347–356 (2005).
37. Moyer, K. E., Kimsey, H. H. & Waldor, M. K. Evidence for a rolling-circle mechanism of phage DNA synthesis from both replicative and integrated forms of CTXφ. *Mol. Microbiol.* **41**, 311–323 (2001).
38. Davis, B. M., Kimsey, H. H., Chang, W. & Waldor, M. K. The *Vibrio cholerae* O139 Calcutta bacteriophage CTXφ is infectious and encodes a novel repressor. *J. Bacteriol.* **181**, 6779–6787 (1999).
39. Das, B., Kumar Ghosh, R., Sharma, C., Vasin, N. & Ghosh, A. Tandem repeats of cholera toxin gene in *Vibrio cholerae* 0139. *The Lancet* **342**, 1173–1174 (1993).
40. Bhuiyan, N. A. *et al*. Changing genotypes of cholera toxin (CT) of *Vibrio cholerae* O139 in Bangladesh and description of three new CT genotypes. *FEMS Immunol. Med. Microbiol.* **57**, 136–141 (2009).

41. Kim, E. J., Lee, C. H., Nair, G. B. & Kim, D. W. Whole-genome sequence comparisons reveal the evolution of *Vibrio cholerae* O1. *Trends Microbiol.* **23**, 479–489 (2015).
42. Kimsey, H. H., Nair, G. B., Ghosh, A. & Waldor, M. K. Diverse CTXΦs and evolution of new pathogenic *Vibrio cholerae*. *The Lancet* **352**, 457–458 (1998).
43. Klinzing, D. C. *et al*. Hybrid *Vibrio cholerae* El Tor lacking SXT identified as the cause of a cholera outbreak in the Philippines. *mBio* **6**, e00047–15, https://doi.org/10.1128/mBio.00047-15 (2015).
44. Waldor, M. K., Tschäpe, H. & Mekalanos, J. J. A new type of conjugative transposon encodes resistance to sulfamethoxazole, trimethoprim, and streptomycin in *Vibrio cholerae* O139. *J. Bacteriol.* **178**, 4157–4165 (1996).
45. Murphy, R. A. & Boyd, E. F. Three pathogenicity islands of *Vibrio cholerae* can excise from the chromosome and form circular intermediates. *J. Bacteriol.* **190**, 636–647 (2008).
46. Grim, C. J. *et al*. Occurrence of the *Vibrio cholerae* seventh pandemic VSP-I island and a new variant. *OMICS J. Integr. Biol.* **14**, 1–7 (2010).
47. Rowe-Magnus, D. A., Guerout, A.-M. & Mazel, D. Bacterial resistance evolution by recruitment of super-integron gene cassettes. *Mol. Microbiol.* **43**, 1657–1669 (2002).
48. Zhou, Y. *et al*. Accumulation of mutations in DNA gyrase and topoisomerase IV genes contributes to fluoroquinolone resistance in *Vibrio cholerae* O139 strains. *Int. J. Antimicrob. Agents* **42**, 72–75 (2013).
49. Heidelberg, J. F. *et al*. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406**, 477–483 (2000).
50. Hu, D. *et al*. Origins of the current seventh cholera pandemic. *Proc. Natl. Acad. Sci.* **113**, E7730–E7739, https://doi.org/10.1073/pnas.1608732113 (2016).
51. Chin, C.-S. *et al*. The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).
52. Feng, L. *et al*. A recalibrated molecular clock and independent origins for the cholera pandemic clones. *PLoS ONE* **3**, e4053, https://doi.org/10.1371/journal.pone.0004053 (2009).
53. Pérez Chaparro, P. J. *et al*. Whole genome sequencing of environmental *Vibrio cholerae* O1 from 10 nanograms of DNA using short reads. *J. Microbiol. Methods* **87**, 208–212 (2011).
54. Chowdhury, F. *et al*. Concomitant enterotoxigenic *Escherichia coli* infection induces increased immune responses to *Vibrio cholerae* O1 antigens in patients with cholera in Bangladesh. *Infect. Immun.* **78**, 2117–2124 (2010).
55. Qadri, F., Svennerholm, A.-M., Faruque, A. S. G. & Sack, R. B. Enterotoxigenic *Escherichia coli* in developing countries: Epidemiology, microbiology, clinical features, treatment, and prevention. *Clin. Microbiol. Rev.* **18**, 465–483 (2005).
56. Faruque, A., Mahalanabis, D., Islam, A. & Hoque, S. Severity of cholera during concurrent infections with other enteric pathogens. *J. Diarrhoeal Dis. Res.* **12**, 214–218 (1994).
57. Davis, B. M., Moyer, K. E., Boyd, E. F. & Waldor, M. K. CTX prophages in classical biotype *Vibrio cholerae*: Functional phage genes but dysfunctional phage genomes. *J. Bacteriol.* **182**, 6992–6998 (2000).
58. Davies, B. W., Bogard, R. W., Young, T. S. & Mekalanos, J. J. Coordinated regulation of accessory genetic elements produces cyclic di-nucleotides for *V. cholerae* virulence. *Cell* **149**, 358–370 (2012).
59. Butler, S. M. *et al*. Cholera stool bacteria repress chemotaxis to increase infectivity. *Mol. Microbiol.* **60**, 417–426 (2006).
60. Chin, C.-S. *et al*. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* **10**, 563–569 (2013).
61. Koren, S. *et al*. Canu: scalable and accurate long-read assembly *via* adaptive k-mer weighting and repeat separation. *Genome Res.* **27**, 722–736, https://doi.org/10.1101/gr.215087.116 (2017).
62. Hunt, M. *et al*. Circlator: automated circularization of genome assemblies using long sequencing reads. *Genome Biol.* **16**, 294 (2015).
63. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
64. O'Leary, N. A. *et al*. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745, https://doi.org/10.1093/nar/gkv1189 (2016).
65. Walker, B. J. *et al*. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
66. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
67. Carver, T. J. *et al*. ACT: the Artemis comparison tool. *Bioinformatics* **21**, 3422–3423 (2005).
68. Li, H. *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
69. Harris, S. R. *et al*. Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327**, 469–474 (2010).
70. Croucher, N. J. *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15, https://doi.org/10.1093/nar/gku1196 (2015).
71. Page, A. J. *et al*. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb. Genomics* **2**, e000083, https://doi.org/10.1099/mgen.0.000083 (2016).
72. Page, A. J. *et al*. Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691–3693 (2015).
73. Capella-Gutiérrez, S., Silla-Martínez, J. M. & Gabaldón, T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**, 1972–1973 (2009).
74. Page, A. J. *et al*. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2**, e000056, https://doi.org/10.1099/mgen.0.000056 (2016).
75. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
76. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.* **32**, 268–274 (2015).
77. Lewis, P. O. A likelihood approach to estimating phylogeny from discrete morphological character data. *Syst. Biol.* **50**, 913–925 (2001).
78. Guindon, S. *et al*. New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Syst. Biol.* **59**, 307–321 (2010).
79. Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2: Improving the ultrafast bootstrap approximation. *Mol. Biol. Evol.* **35**, 518–522 (2018).
80. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242–W245, https://doi.org/10.1093/nar/gkw290 (2016).
81. Zankari, E. *et al*. Identification of acquired antimicrobial resistance genes. *J. Antimicrob. Chemother.* **67**, 2640–2644 (2012).

Acknowledgements

We acknowledge the support of the dedicated field and laboratory workers at the icddr,b involved in this study, and thank Derek Pickard for technical help with DNA extractions. We also thank the sequencing and Pathogen Informatics teams at the Wellcome Sanger Institute for help with processing samples and depositing sequencing data. This work was supported by Wellcome (grants 098051, 206194) and by the International Centre for Diarrhoeal Disease Research, Bangladesh (icddr,b). This study was supported by grants from the National Institutes of Health, including grants from the National Institute of Allergy and Infectious Diseases

(AI106878 [F.Q.], AI058935 [F.Q.], and AI103055 [F.Q.]). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript. icddr,b is grateful to the Governments of Bangladesh, Canada, Sweden and UK for providing core/unrestricted support. M.J.D. is supported by a Wellcome Sanger Institute PhD Studentship.

Author Contributions

N.R.T. and F.Q. designed and supervised the study. S.S. and Y.A.B. co-ordinated and performed experimental work. M.J.D., D.D. and M.I.U. analysed the data. M.J.D., D.D. and M.I.U. wrote the manuscript, with major contributions from M.H.A., F.Q. and N.R.T. All authors contributed to the editing of the manuscript.

Additional Information

Supplementary information accompanies this paper at https://doi.org/10.1038/s41598-019-41883-x.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.