nature microbiology

Article

https://doi.org/10.1038/s41564-023-01472-1

Genomic epidemiology reveals multidrug resistant plasmid spread between *Vibrio cholerae* lineages in Yemen

Received: 7 July 2022

Accepted: 11 August 2023

Published online: 28 September 2023

Check for updates

Florent Lassalle[1] ✉, Salah Al-Shalali[2], Mukhtar Al-Hakimi[2], Elisabeth Njamkepo[3], Ismail Mahat Bashir[4], Matthew J. Dorman[1,5], Jean Rauzier[3], Grace A. Blackwell[1,6], Alyce Taylor-Brown[1], Mathew A. Beale[1], Adrián Cazares[1], Ali Abdullah Al-Somainy[7], Anas Al-Mahbashi[2], Khaled Almoayed[7], Mohammed Aldawla[8], Abdulelah Al-Harazi[7], Marie-Laure Quilici[3,11], François-Xavier Weill[3,11], Ghulam Dhabaan[9,11] ✉ & Nicholas R. Thomson[1,10,11] ✉

Since 2016, Yemen has been experiencing the largest cholera outbreak in modern history. Multidrug resistance (MDR) emerged among *Vibrio cholerae* isolates from cholera patients in 2018. Here, to characterize circulating genotypes, we analysed 260 isolates sampled in Yemen between 2018 and 2019. Eighty-four percent of *V. cholerae* isolates were serogroup O1 belonging to the seventh pandemic El Tor (7PET) lineage, sub-lineage T13, whereas 16% were non-toxigenic, from divergent non-7PET lineages. Treatment of severe cholera with macrolides between 2016 and 2019 coincided with the emergence and dominance of T13 subclones carrying an incompatibility type C (IncC) plasmid harbouring an MDR pseudo-compound transposon. MDR plasmid detection also in endemic non-7PET *V. cholerae* lineages suggested genetic exchange with 7PET epidemic strains. Stable co-occurrence of the IncC plasmid with the SXT family of integrative and conjugative element in the 7PET background has major implications for cholera control, highlighting the importance of genomic epidemiological surveillance to limit MDR spread.

Since 2016, Yemen has seen the largest epidemic of cholera ever recorded. This occurred against the backdrop of a civil war turned international conflict and famine, which together fuelled extensive population movement, with more than four million people internally displaced by the end of 2020[1]. The Electronic Disease Early Warning System, a surveillance programme coordinated by the Ministry of Public Health and Population of Yemen (MPHP) in Sana'a tasked with monitoring the epidemic[2], recorded a total of almost 2.4 million suspected cholera cases up until August 2019[3]. These cases exhibited a seasonal profile, with peaks in July 2017 and September 2018 (16,000 and 50,000 cases per week, respectively)[3]. The lower reported case incidence in 2018 was ascribed to the mass vaccination campaign led

[1]Parasites and Microbes Programme, Wellcome Sanger Institute, Hinxton, UK. [2]Faculty of Science, Sana'a University, Sana'a, Yemen. [3]Institut Pasteur, Université Paris Cité, Unité des Bactéries pathogènes entériques, Paris, France. [4]WHO Yemen country office, Sana'a, Yemen. [5]Churchill College, Cambridge, UK. [6]EMBL-EBI, Hinxton, UK. [7]National Centre of Public Health Laboratories, Sana'a, Yemen. [8]Ministry of Public Health, Infection Control Unit, Sana'a, Yemen. [9]Department of Laboratory Medicine and Pathobiology, University of Toronto, Toronto, Ontario, Canada. [10]London School of Hygiene and Tropical Medicine, London, UK. [11]These authors jointly supervised this work: Marie-Laure Quilici, François-Xavier Weill, Ghulam Dhabaan, Nicholas R. Thomson. ✉e-mail: florent.lassalle@sanger.ac.uk; ghulam.dhabaan@utoronto.ca; nrt@sanger.ac.uk

by the World Health Organization and United Nation Children's Fund, who delivered the oral cholera vaccine to 540,000 people in August 2018 (and 387,000 at follow-up in September) in targeted districts in Aden, Hudaydah and Ibb governorates[4,5]. Despite this mass vaccination campaign, cholera cases were recorded nationwide in 2019, peaking at over 30,000 cases per week and case numbers declined at a slower rate than in previous years[3].

Pandemic cholera is caused by specific phylogenetic lineages of the bacterium *Vibrio cholerae* that are associated with epidemic spread, and which carry lipopolysaccharide O-antigens of serogroups O1 or O139. The large majority of epidemic strains associated with cholera outbreaks from the past 60 years belong to the seventh pandemic El Tor (7PET) lineage of *V. cholerae* O1, which has spread globally in three pandemic waves[6]. We previously used genomic epidemiology to show that the first two waves of the cholera outbreak in Yemen (2016 and 2017) were driven by a single clonal expansion[7] belonging to Wave 3 of the global 7PET lineage and had an Ogawa serotype. This indicated the Yemen outbreak was seeded by a single international transmission event linked to the 7PET sub-lineage denoted T13 (ref. 7).

Ongoing surveillance activities in Yemen found that the fluctuating peaks in incidence were accompanied by a sudden change in the antibiotic susceptibility profile reported by the reference laboratory at the MPHP in Sana'a. Although strains isolated in 2016–2018 were sensitive to most of the antibiotics usually used for the treatment of cholera (except quinolones, where reduced susceptibility to ciprofloxacin prevented the use of this antibiotic as a single-dose treatment), by 2019, resistance was observed for multiple drugs including third-generation cephalosporins, macrolides (including azithromycin) and co-trimoxazole. Although the main treatment for cholera is rehydration therapy, antibiotics can be used to limit the volume and duration of the acute watery diarrhoea, and reduce the risk of transmission[8–10]. In Yemen, macrolides were used extensively up to early 2019 to treat moderate to severe cases of cholera in pregnant women and children, the latter forming the large majority of cases[11]. Multiple drug resistance (MDR) in *V. cholerae* is strongly associated with the acquisition of mobile genetic elements (MGEs) such as SXT family integrative and conjugative elements (SXT ICE) or plasmids of the incompatibility type C (IncC; formerly known as IncA/C_2)[12], which often carry and disseminate antimicrobial resistance (AMR) gene cargo[13].

We hypothesized that the MDR phenotype seen in the Yemen *V. cholerae* isolates from 2019 could be explained either by gain of resistance (through de novo mutations or acquisition of a resistance-conferring MGE) in the previously susceptible 7PET-T13 *V. cholerae* strain already circulating in Yemen, or through the replacement of that strain with locally or globally derived MDR strain(s). Distinguishing between these hypotheses is important for understanding the ongoing dynamics of cholera in Yemen and will be important for cholera control strategies. We therefore applied genomic epidemiology approaches to determine the molecular basis for the observed switch to the MDR phenotype and its link to evolutionary dynamics of pandemic cholera. In doing so, we highlight the role of globally circulating MGEs in making an epidemic pathogen resistant to multiple drugs and subsequently reducing treatment options. We also show that these MGEs and their cargo AMR genes were repeatedly exchanged among diverse *V. cholerae* lineages found in Yemen.

Results

V. cholerae in Yemen in 2018 and 2019

The National Centre of Public Health Laboratories (NCPHL) in Sana'a, the capital city of Yemen, received 6,311 and 3,225 clinical samples from suspected cholera patients in 2018 and 2019, respectively (Supplementary Table 1). Of these, 2,204 (35%) and 2,171 (67%) were confirmed to be positive for *V. cholerae* O1 by culture (identification based on biochemical tests and detection of Ogawa and Inaba serotypes) (Supplementary Table 1 and Extended Data Fig. 1). Among the 1,642 *V. cholerae* isolated at the NCPHL from January to October 2018, 623 were tested for susceptibility to a range of antibiotics using the disk diffusion method, of these 620 (99.6%) were phenotypically resistant to nalidixic acid and nitrofurantoin, but otherwise sensitive to all other antimicrobials tested (Extended Data Fig. 2 and Supplementary Table 1). By contrast, all tested *V. cholerae* isolates (n = 2,172) from January 2019 onwards were resistant to nalidixic acid, azithromycin, co-trimoxazole and cefotaxime (Supplementary Fig. 2 and Supplementary Table 1), a pattern maintained up to late 2021 (see ref. 14 for 2020–2021 cholera surveillance data). The transition in phenotype occurred during November 2018, when 159/175 (90.8%) tested isolates already showed the MDR profile. Of the 2018–2019 clinical *V. cholerae* isolates, 250 were randomly chosen for further characterization (Supplementary Table 2). These samples originated from 8 of the 21 Yemen governorates, comprising 71 of 333 districts, with 101 samples collected in 2018 (from mid-July to late October) and 149 collected in 2019 (from late February to late April and from early August to mid-October). In addition, ten environmentally derived strains were isolated from sewage in Sana'a in October 2019 (Supplementary Table 2). Extended antibiotic sensitivity testing of these 260 isolates at NCPHL and of a subset (n = 22) at Institut Pasteur (IP) (Extended Data Fig. 3) confirmed the phenotypic switch to MDR observed in the wider sample set, further showing that all tested 2019 strains were resistant to ampicillin, cefotaxime, nalidixic acid, azithromycin, erythromycin and co-trimoxazole (Supplementary Text).

Phylogenetic diversity of *V. cholerae* in Yemen (2018, 2019)

We isolated a single colony for 240 of the 260 *V. cholerae* isolates indicated above, and multiple independent colony picks for the remaining 20, for a total of 281 isolates on which we performed whole-genome sequencing (Extended Data Fig. 3 and Supplementary Tables 2 and 3). After quality filtering, 232 high-quality isolate genomes were assembled (selecting a single isolate from each initial sample) (Supplementary Table 4), which we combined with 650 previously published *V. cholerae* O1 and non-O1 genomes for context (Supplementary Table 5 and Extended Data Fig. 3). We inferred a core-genome phylogeny for this genome set (n = 882), which described the sequenced diversity of the *V. cholerae* species, rooted by the genomes that belong to its newly described sister species *V. paracholerae*[15]. We subdivided *V. cholerae* genomes into 11 clades, referred to henceforth as *Vc*A to *Vc*K (Fig. 1 and Supplementary Table 5). *Vc*H contained all 7PET epidemic lineage genomes utilized in this dataset, including the majority (216/232) of the Yemen 2018–2019 genomes and all 42 previously reported 2016–2017 Yemeni genomes[7] (Extended Data Fig. 4).

Although Yemeni *Vc*H isolates show limited genomic diversity (99.98%–100.00% average nucleotide identity (ANI) similarity; 0 to 97 single nucleotide polymorphisms (SNPs)), the remaining 16 Yemeni genomes belonged to clades *V. paracholerae* (*Vpc*), *Vc*D and *Vc*K, and were overall more diverse than *Vc*H isolate genomes (96.24%–99.99% ANI similarity) (Fig. 1 and Table 1); these represent 'non-7PET' lineages. Among these, we found five distinct sequence types in three lineages: *Vpc* (n = 1; ST1499), *Vc*D (n = 21; ST555, ST1020 and ST1498) (Supplementary Table 5) and *Vc*K (n = 2; ST170) (Fig. 1). Collectively, these non-7PET isolates comprised 8% of the clinical isolates (21/254) and 30% of environmentally derived isolates (3/10).

Although highly clonal, the phylogenetic structure within the *Vc*H clade allowed it to be further subdivided into subclades *Vc*H.1 to *Vc*H.10 (Extended Data Fig. 5). All 216 *Vc*H Yemen 2016–2019 isolates fell within *Vc*H.9, which corresponds to the T13 sub-lineage of 7PET Wave 3 (ref. 7). We selected one representative isolate (CNRVC190243) of *Vc*H.9 and used PacBio sequencing to generate long reads in addition to the Illumina short reads already obtained, which enabled us to generate a closed hybrid assembly (Supplementary Text). To obtain greater phylogenetic resolution within *Vc*H.9, we mapped *Vc*H.9 read sets to our new *Vc*H.9 CNRVC190243 reference genome to produce a multiple whole-genome alignment. To capture

https://doi.org/10.1038/s41564-023-01472-1

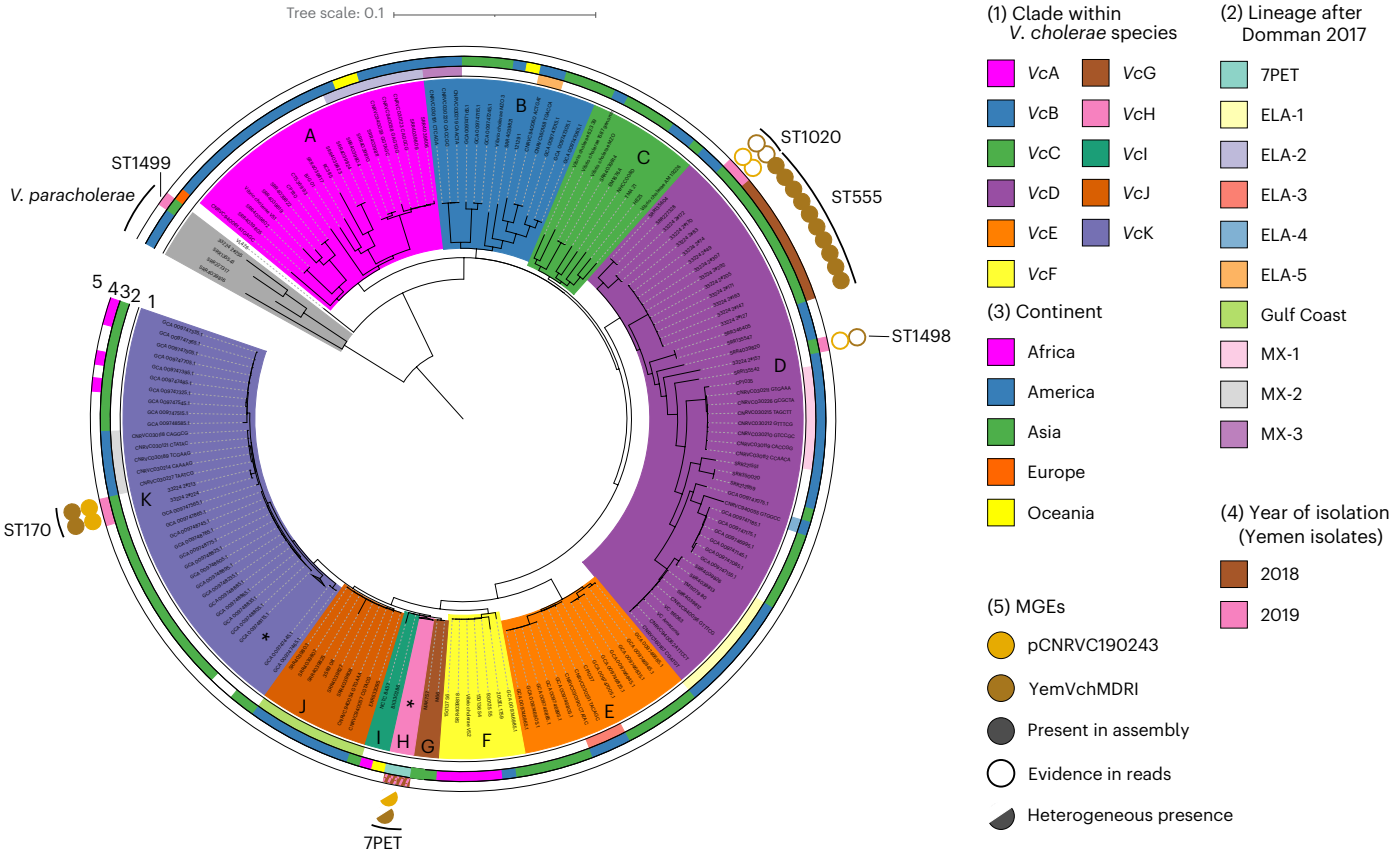


Fig. 1 | Phylogenetic diversity of *V. cholerae* isolates from Yemen. ML phylogeny of 882 assembled *V. cholerae* genomes based on the 37,170 SNP sites from the concatenated alignments of 291 core genes. Low-diversity clades (*Vc*H and part of *Vc*K) are collapsed and marked by black stars. Clades are highlighted with background colours (legend key 1). Coloured rings outside the tree depict the match with previously described lineages (ring 2), the geographical origin of isolates at the level of continents (ring 3) and their year of isolation when from Yemen (ring 4). The presence of parts of the plasmid pCNRVC190243 are indicated by coloured circles (ring 5 in A): IncC plasmid backbone (light brown) and the MDR PCT Yem*Vch*MDRI (dark brown); full circles indicate more than 70% coverage in assemblies of the reference length, hollow circles indicate 30%–70% coverage in assemblies and confirmed presence based on mapped reads, with even coverage over the MGE reference sequence, whereas half-circles represent heterogeneous presence in a collapsed clade. The scale bar represents the number of nucleotide substitutions per site.

Table 1 | Number of *V. cholerae* isolate genomes from Yemen by year and phylogenetic lineage

Year	Total	Clades				Clusters				Not determined[d]
		Non-7PET			7PET	H.9.e	H.9.f	H.9.g	H.9.h	
		Vpc[a]	*VcD*[b]	*VcK*[a]	*VcH/H.9*[c]					
2016[e]	8				8	7	1			
2017[e]	34				34	29	5			
2018	112		17		87	3		6	78	8
2019	169	1	4	2	151			150	1	11
Total	323	1	21	2	280	39	6	156	79	19

[a]Assigned based on the '882 assembled *V. cholerae* genomes' dataset. [b]Assigned based on the '33 mapped *Vc*D genomes' dataset. [c]Assigned based on the '456 mapped 7PET genomes' dataset. [d]Poor quality genome data or no coverage of the bacterial genomes; for example, in case of complete contamination by ICP1 virus genome. [e]As reported in ref. 7.

minute details of the diversification of *Vc*H.9 during the outbreak, we used all 2018–2019 Yemeni isolate genomes available to us: we included seven genomes for which mapping quality was satisfactory although assembly had failed quality control, and 15 genomes for duplicate isolates independently cultured at IP (Extended Data Fig. 3), for a total of 238 *Vc*H.9 sequences. Our final alignment and resultant 'mapped genome tree' included another 218 previously published genomes that reside in this subclade and close outgroups, for a total of 456 genomes (Supplementary Table 6). This approach allowed us to further subdivide *Vc*H.9 into phylogenetic clusters named *Vc*H.9.a to *Vc*H.9.h. (Fig. 2a).

Yemeni 7PET *V. cholerae* genomes form a monophyletic group (clusters *Vc*H.9.e to *Vc*H.9.h; Table 1), emerging from the genetic diversity of East African genomes (clusters *Vc*H.9.c and *Vc*H.9.d), which in turn branch out of a cluster of South Asian genomes (*Vc*H.9.b), consistent with previous observations on the origins of 7PET-T13, introduced from South Asia into Africa[7,16]. Clusters *Vc*H.9.g and *Vc*H.9.h comprise the majority of 2018–2019 Yemen isolates (235/281) and

form a well-supported clade (94% bootstrap) that branches from within *Vc*H.9.f.

Spatio-temporal distribution of *V. cholerae* isolates

To delineate the evolutionary dynamics of the cholera outbreak in Yemen, we plotted *Vc*H.9 isolates by phylogenetic cluster over time (based on the date of sample collection) and between administrative divisions (linked to reporting hospital). From Fig. 2b, it is clear that each annual wave was dominated by a single cluster: 2016 and 2017 by *Vc*H.9.e, 2018 by *Vc*H.9.h and 2019 by *Vc*H.9.g. There was no evidence of geographic restriction for any of these clusters, even when accounting for dispersal over time (Fig. 2c,d, Supplementary Table 6 and online Supplementary data at https://doi.org/10.6084/m9.figshare.19097111). There was also no significant correlation between spatial and temporal distances, or between the spatial and phylogenetic distances (Supplementary Table 7).

However, these data did show a positive correlation between the temporal and phylogenetic distances ($R^2 = 0.181$; Mantel test $P < 10^{-6}$) (Supplementary Table 7), with root-to-tip distances significantly correlated with sampling date (Pearson's $R^2 = 0.437$; $P < 10^{-15}$).

We inferred a recombination-free, timed phylogeny for *Vc*H.9 using a Bayesian framework (Extended Data Fig. 6), which revealed that the most recent common ancestor (MRCA) of all Yemeni *V. cholerae* 7PET-T13 genomes was estimated to have existed in February 2015 (95% confidence interval, April 2014 and July 2015). Moreover, the MRCAs for clusters *Vc*H.9.e and *Vc*H.9.f (mostly sampled in 2016 and 2017) were dated May and June 2015 respectively, and the MRCAs for clusters *Vc*H.9.g and *Vc*H.9.h (sampled in 2018 and 2019) were dated February and March 2017 respectively. In addition, we dated the MRCA of the clade grouping clusters *Vc*H.9.g and *Vc*H.9.h, which represent the majority of 2018–2019 Yemen isolates, to September 2016 (Extended Data Fig. 6).

In contrast to 7PET isolates, the distribution of non-7PET isolates (clades *Vc*D, *Vc*K and *Vpc*) across Yemen was mostly sporadic. Although we found some of the non-7PET isolates were closely related and occurred in close spatio-temporal range (Supplementary Table 5 and Supplementary data at https://doi.org/10.6084/m9.figshare.19097111), we found no evidence of long-range spread of the non-7PET isolates across Yemen (Supplementary Text), a pattern that thus remains characteristic of 7PET *V. cholerae* isolates linked to epidemic disease.

Predicted phenotypic properties of *V. cholerae* isolates

Consistent with our previous report[7], Yemeni *Vc*H.9 isolates—which all belong to 7PET-T13 sub-lineage—all carried genes or mutations known to confer resistance to trimethoprim (*dfrA1*), to nalidixic acid (*gyrA*_S83I and *parC*_S85L) and to nitrofurans (*nfsA*_R169C and *nfsB*_Q5Stop). They also carried the *Vibrio* pathogenicity island 1 (VPI-1)—encoding the toxin co-regulated pilus—and VPI-2, the *Vibrio* seventh pandemic islands I and II (VSP-I and VSP-II), and the CTX prophage, which all featured the cholera toxin genes, *ctxAB*, of the allelic type *ctxB7*. None of the non-7PET genomes from Yemen possessed a CTX phage or the *ctxAB* genes.

All Yemeni 2018–2019 *Vc*H isolates were predicted to be the O1 serogroup (except for three isolates for which genomic data were insufficient) (Supplementary Text and Extended Data Fig. 7) and were predicted to be Ogawa serotype, except two that showed a disruption in *wbeT*, indicative of an Inaba phenotype (YE-NCPHL-18053 and YE-NCPHL-19014, with gene truncation and point mutation respectively) (Supplementary Table 6).

Genome variation of *Vc*H.9 (7PET-T13) isolates in Yemen

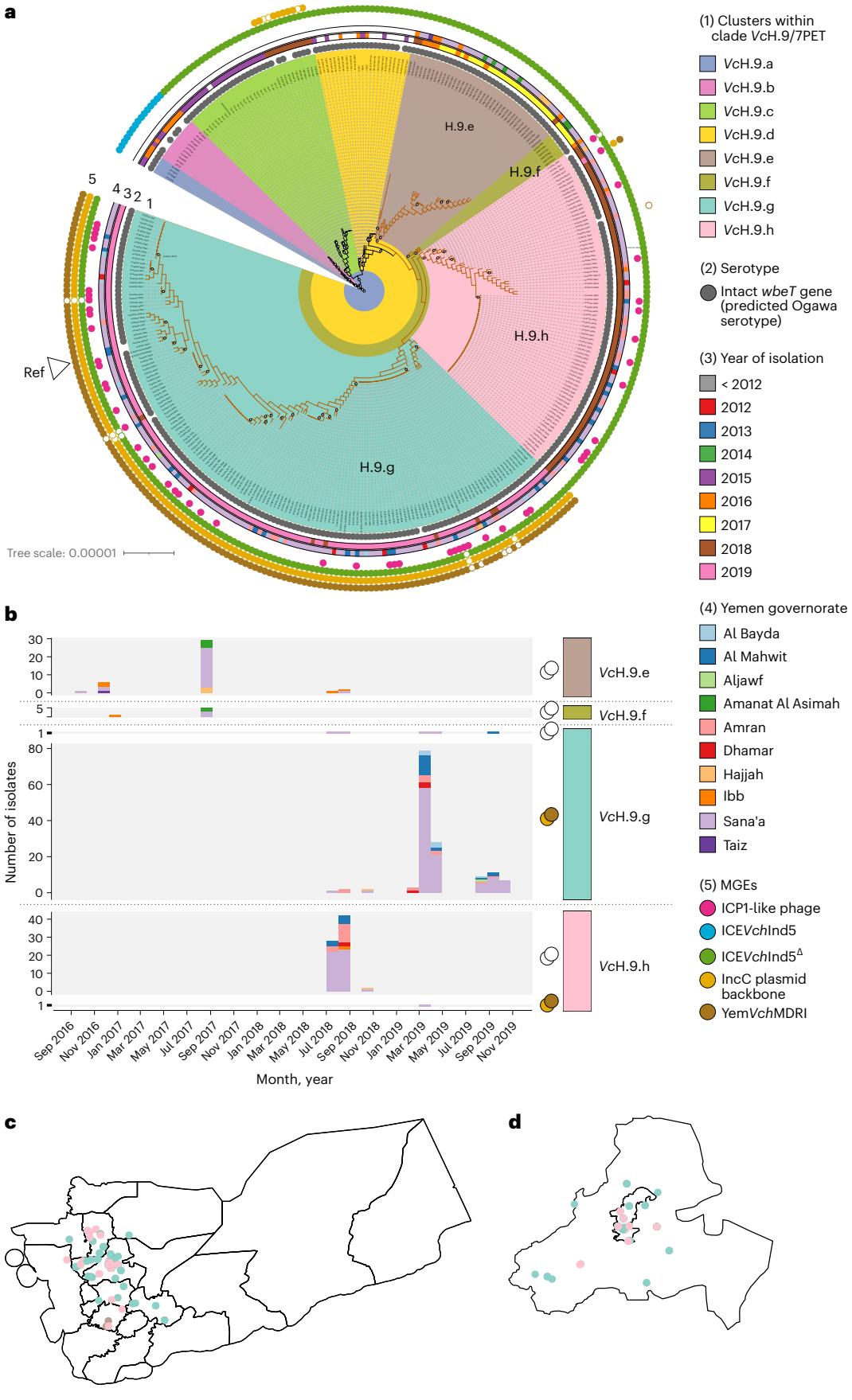
Given the change in antimicrobial susceptibility seen in the 2018–2019 Yemen isolates, we compared in detail the *Vc*H.9 isolate genomes from Yemen with each other and with related isolates taken elsewhere. We identified 3, 4 and 21 fixed SNPs in the crown clade containing *Vc*H.9.e,f,g,h, the clade containing *Vc*H.9.g,h, and *Vc*H.9.h, respectively (including 2, 2 and 11 non-synonymous SNPs, respectively) (Supplementary Table 8). Changes fell largely within genes predicted to be involved in carbohydrate metabolism, signal transduction and chemotaxis, none of which could be directly linked to change in virulence (Supplementary Table 8).

Previously, the 2016–2017 Yemeni isolates carried an SXT ICE differing by only three or four SNPs from the ICE*Vch*Ind5/ICE*Vch*Ban5 reference sequence (GenBank accession GQ463142.1)[17], but which possessed a 10 kb deletion in variable region III, explaining the phenotypic loss of resistance to streptomycin, chloramphenicol and sulphonamides (only retaining resistance to trimethoprim via the *dfrA1* gene)[7]. All 2018–2019 *Vc*H.9 genomes carried the same SXT ICE deletion variant (Supplementary Table 6), showing a maximum of two pairwise SNP differences and indicating the change in AMR profile was not linked to variation in SXT ICE.

Looking across all genes within the pangenome, the only variation directly associated with the Yemen 2018–2019 genomes, compared with those sequenced in the period 2016–2017, was the presence of a 139 kb plasmid, which we named pCNRVC190243 (Supplementary Table 9). The backbone of this plasmid includes a replicon of the IncC type (previously known as IncA/C_2 subtype[12]), as well as genes encoding a complete type F conjugative apparatus and a mobility region of the family MOB_H, suggesting it is self-transmissible. Plasmid pCNRVC190243 also carries a 20 kb genomic region (which we denoted Yem*Vch*MDRI); this is a pseudo-compound transposon (PCT)—a structure bounded by IS*26* elements[18]—and includes a class 1 integron with *aadA2* encoding resistance to streptomycin and spectinomycin as a gene cassette, associated with an IS*CR1* element carrying the extended spectrum beta-lactamase bla_{PER-7} gene, a structure similar to one previously seen in *Acinetobacter baumanii*[19,20]. It is also predicted to encode a quaternary ammonium compound efflux pump (*qac*), sulphonamide resistance (*sul1*) and macrolide resistance (*mph*(A), *mph*(E) and *msr*(E)) (Fig. 3). We found that

Fig. 2 | Phylogenetic diversity and spatio-temporal distribution of *V. cholerae* 7PET-T13 isolates (*Vc*H.9) from Yemen. a, Subtree of the ML phylogeny of 456 7PET genomes mapped to reference *Vc*H.9 strain CNRVC190243 genome, including 335/456 genomes covering *Vc*H.9 (as defined in Supplementary Fig. 5), which corresponds to the 7PET-T13 sub-lineage and close South Asian relatives. The full tree containing the 456 genomes is available as supplementary material on figshare (https://doi.org/10.6084/m9.figshare.16595999) and was obtained based on 2,092 SNP sites from concatenated whole-chromosome alignments. Brown branches indicate the clade grouping all Yemeni 7PET-T13 isolates. Bootstrap support greater than 70% is indicated by white circles. Phylogenetic clusters within *Vc*H.9 are highlighted with background colours (legend key 1). Coded tracks outside the tree depict the serotype of isolates (ring 2) as predicted from genomic data, year of isolation when isolated in 2012 or later (ring 3) and the governorate of isolation if in Yemen (ring 4). The presence of MGEs is indicated by coloured circles in the outermost track (ring 5): ICP1-like phage (pink), SXT ICE ICE*Vch*Ind5 (blue), ICE*Vch*Ind5^A that is featuring the characteristic 10-kb deletion in the variable region III (green), IncC plasmid backbone (light brown) and the MDR PCT Yem*Vch*MDRI (dark brown); filled and unfilled circles indicate different levels of coverage in assemblies (as in Fig. 1 legend). The position of the reference sequence to which all other genomes were mapped to generate the alignment is labelled. The scale bar represents the number of nucleotide substitutions per site. **b**, Frequency of each phylogenetic subcluster among Yemen isolates per month since the onset of the Yemen outbreak. Where relevant, the cluster group is subdivided by the presence or absence of the IncC plasmid as indicated by the filled brown (present) or open (absence) circle on the right of the chart. The contribution of each governorate of isolation is indicated by the coloured portion of each bar. **c**,**d**, A map of Yemen governorates (**c**) and a focus on the Sana'a and Amanat Al Asimah governorates (inner and outer capital city; **d**), with dots corresponding to isolates, coloured by phylogenetic subcluster.

https://doi.org/10.1038/s41564-023-01472-1



https://doi.org/10.1038/s41564-023-01472-1

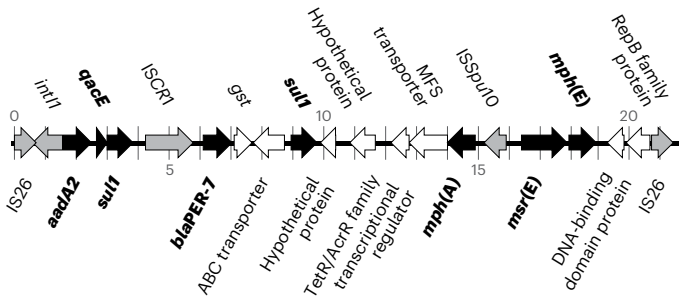


Fig. 3 | Genetic organization of the MDR PCT Yem*Vch*MDRI. AMR genes are filled in black and labelled in bold; genes encoding endonucleases transposases and other genes involved in genetic mobility are filled in grey. Genomic position is indicated by tick marks every kilobase, in reference to the pCNRVC190243 plasmid coordinates.

pCNRVC190243 was present in 6/89 (6.7%) Yemeni *Vc*H.9 isolates from 2018, but this rose to 100% (151/151) in 2019 (Fig. 2b). This was linked to a specific phylogenetic cluster: only 1/79 (1.3%) *Vc*H.9.h isolates harboured the plasmid, compared with all (156/156) *Vc*H.9.g isolates (Fig. 2a).

Distribution and relatedness of MDR MGEs

Analysis of the broader phylogenetic context of pCNRVC190243 and associated Yem*Vch*MDRI showed that the plasmid was also present in three *Vc*D (ST1499 and ST1020) and two *Vc*K (ST170) isolates collected in 2019 in Yemen. Comparing the full-length sequence of all pCNRVC190243 plasmids from *Vc*H.9, *Vc*K and *Vc*D isolates showed that all sequences were identical except for two isolates: one varied by a single SNP resulting in an amino acid change S71F in the sulphonamide resistance protein Sul1 (YE-NCPHL-19105; G26720A SNP); the other by a single intergenic SNP.

We also found the Yem*Vch*MDRI element integrated into chromosome 2, without the pCNRVC190243 backbone, in all 18 of the ST555 isolates (Fig. 1 and Supplementary Text). More broadly, by searching the public genome databases (Supplementary Table 10), we found related but non-identical elements in other *V. cholerae*: an IncC plasmid, named pYA00120881 (GenBank accession MT151380), was identified in 13 closely related *Vc*H.9.a and *Vc*H.9.c isolates (Fig. 2a) that were collected in 2018 in Zimbabwe[16]. The backbones of pCNRVC190243 and pYA00120881 share 99.98% nucleotide sequence identity, but pYA00120881 carries a different MDR genomic region—featuring a *bla* gene encoding a CTX-M-15 extended spectrum beta-lactamase—inserted at the same locus (Extended Data Fig. 8). Furthermore, 59 *V. cholerae* O139 (ST69) isolates collected in China from 1998 to 2009 (publicly available genomic data released in BioProject PRJNA303115)[21,22] carry IncC-type plasmids that show similarity to pCNRVC190243 and also include Yem*Vch*MDRI-like PCT elements, albeit lacking IS*CR*1 and the bla_{PER-7} gene.

Importantly, when using the Yem*Vch*MDRI sequence alone to search the database (Supplementary Table 11), we found that the genome of *V. cholerae* ST555 strain 338360 (Supplementary Table 5) shared 100% nucleotide identity with the Yem*Vch*MDRI carried by Yemeni ST555 genomes, including the bla_{PER-7}-carrying IS*CR*1 (IS*CR*1$_{bla PER-7}$) (Supplementary Table 11). Likewise, IS*CR*1$_{bla PER-7}$ has also been observed previously in the genomes of *A. baumanii* strains[19,23] from France and the United Arab Emirates. Those from United Arab Emirates were located on the plasmid pAB154, where the sequence homology with IS*CR*1$_{bla PER-7}$ extended beyond the canonical element and included Yem*Vch*MDRI flanking regions, suggesting that the IS*CR*1$_{bla PER-7}$ carried by pAB154 is derived from Yem*Vch*MDRI, or a closely related element (Extended Data Fig. 9). What is more, outside *V. cholerae*, pCNRVC190243- and/or Yem*Vch*MDRI-like elements are widely distributed across other genera, with *Escherichia coli*, *Salmonella enterica* and *Klebsiella pneumoniae* genomes presenting >95% shared nucleotide *k*-mers (Supplementary Tables 10 and 11), with the latter possessing the closest matches outside *V. cholerae*. This indicates that similar regions may be widely distributed in MGEs across bacterial taxa and stably maintained in 7PET genomes.

Discussion

In Yemen, pregnant women and children with cholera (one-third of cholera patients were aged 15 or under)[11] were treated with erythromycin and azithromycin from 2016 until late 2018, at which point there was a sudden change in the observed antimicrobial susceptibility profile in *V. cholerae* isolated from patients. Although strains isolated in 2016–2018 were largely sensitive to antibiotics usually used for cholera treatment (excepting quinolones), by 2019 most isolates were phenotypically resistant to multiple therapeutically relevant drugs, including third-generation cephalosporins and macrolides (including azithromycin). Tetracyclines remained a viable treatment option.

Through our genomic epidemiology analysis we showed that despite significant seasonal fluctuation in incidence, the vast majority of cholera in Yemen was caused by the globally circulating 7PET-T13 lineage (*Vc*H.9) derived from a single introduction. From the inferred phylogeny we were able to subtype Yemeni 7PET-T13 genomes into four different phylogenetic clusters that dominated at different points in time during the outbreak. We observed two large clonal expansions for the sister clades that dominated in 2018 and 2019, both of which first emerged in early 2017. Our data showed that the switch in AMR phenotype coincided with the appearance in late 2018 of plasmid pCNRVC190243 in isolates belonging to the 2019 *Vc*H.9.g phylogenetic cluster. Plasmid pCNRVC190243 carries the PCT Yem*Vch*MDRI, which in turn comprises a type 1 integron and the IS*CR*1$_{bla PER-7}$ element. Yem*Vch*MDRI confers resistance to third-generation cephalosporins, streptomycin (and spectinomycin), macrolides and sulphonamides, plus disinfectant tolerance provided by the *qac* gene[24]. Importantly, pCNRVC190243 (carrying the PCT Yem*Vch*MDRI) was also found in a small number of different non-7PET isolates from Yemen collected in 2018–2019 as well as the only *Vc*H.9.h isolate taken in 2019.

Although the plasmid pCNRVC190243 carrying the PCT Yem*Vch*MDRI appears to be a novel composite element, plasmid pYAM00120881, identified in *Vc*H.9 *V. cholerae* isolates from Zimbabwe in 2018[16], shares an almost identical plasmid backbone, albeit one lacking PCT Yem*Vch*MDRI. Conversely, although rare in the public databases, elements highly similar to Yem*Vch*MDRI occurred in diverse lineages including: *V. cholerae* O139, *V. cholerae* ST555 and two *A. baumanii* strains. Detailed comparison of these PCT-related elements suggests they all are derived from a common ancestral element, and the presence of PCT Yem*Vch*MDRI in multiple ST555 strains isolated from different geographical origins suggests this element is widely distributed and has been acquired multiple times by this and other *V. cholerae* sequence types, and other bacterial genera. This is also consistent with the fact that strain 338360—a ST555 isolate from India—is distinguished from our Yemeni ST555 reference strain CNRVC019247 by 436 SNPs over both chromosomes, indicating that the Yemen isolates, although related and carrying the same PCT, do not share a recent common ancestor with strain 338360. Furthermore, because pCNRVC190243 and Yem*Vch*MDRI can self-mobilize, it is possible that Yem*Vch*MDRI would have transposed onto the progenitor of pCNRVC190243 in the context of the Yemen cholera outbreak, and therefore be subject to antibiotic selection. However, more data would be needed to confirm this.

What is clear is that acquisition of the pCNRVC190243 plasmid containing the Yem*Vch*MDRI element by an ancestor of the Yemeni *Vc*H.9.g isolates was followed by its dramatic spread, a clonal expansion that we show occurred in 2018, a time when the treatment regimen was to treat symptomatic cases with macrolides. It is possible to explain the distribution of pCNRVC190243 by multiple acquisitions of the plasmid from independent sources or, more parsimoniously, as direct horizontal gene transfer events between the epidemic and endemic

V. cholerae strains. The large population sizes attained by the epidemic lineages in Yemen make spillover from the dominant cluster at the time, *Vc*H.9.g, probable. Consistent with this, our microbiology, serotyping and sequencing data showed that four of the samples analysed here contained both ST555 and 7PET strains (Supplementary Text), indicative of a limited number of mixed infections from individual patients.

Recently, it has been shown that the presence of two defence systems, called DdmABC and DdmDE, destabilizes plasmids in *V. cholerae* 7PET lineage isolates. However, large IncC-type plasmids are an exception because the Ddm systems only give a competitive disadvantage to plasmid-bearing cells, which is probably overcome in the presence of selection for the function of the plasmid cargo; for example, antibiotic resistance[25]. Having said that, although other MDR IncC plasmids have been previously observed in *V. cholerae* in Democratic Republic of the Congo, Kenya and Zimbabwe, unlike in Yemen these were linked only to sporadic cases or small-scale cholera outbreaks, despite uncontrolled use of antibiotics, and were not linked to T13 lineage isolates. Apart from the Yemen cholera outbreak, the only other example of a massive clonal expansion of a *V. cholerae* lineage carrying an MDR IncC plasmid was the T13 *Vc*H.9.c clone responsible for the Zimbabwean cholera outbreak of 2018, which lasted six months with over 10,000 suspected cases. This observation was so striking that through detailed comparative analysis of the genomes we showed that T13 isolates uniquely carry a 10 kb deletion in the SXT ICE (ICE*Vch*Ind5/ICE*Vch*Ban5). Because the presence of SXT ICE has been proposed previously to prevent the stable replication of IncC-type plasmids through an unknown functional interference[7,16], it is possible this deletion impairs the putative interference mechanism and allows an SXT ICE and an IncC plasmid to stably propagate together in T13 strains.

The emergence of this MDR pathogen demonstrates the necessity of continued genomic surveillance of the microbial population associated with the ongoing Yemen cholera outbreak, and for new outbreaks that may take place in regionally connected areas. Such surveillance will enable Yemeni public health authorities to rapidly adapt clinical practices to minimize AMR selective pressures. This also warrants increased efforts in research on the molecular mechanisms and evolution of interactions between MGEs, to learn about the constraints ruling their colonization of bacterial genomes. Such knowledge is essential for us to be able to disentangle the role of MGEs from that of their bacterial hosts in driving epidemics, so to propose practical definitions of pathogens that focus on the relevant genes, mobile elements or prokaryotic organisms, and to implement appropriate molecular epidemiology surveillance schemes.

Methods

Definitions and surveillance data

Cholera cases were notified to the MPHP and recoded through the Electronic Disease Early Warning System[2]. Suspected and confirmed cholera cases were defined according to the World Health Organization in a declared outbreak setting. Briefly, a suspected case is any person presenting with or dying from acute watery diarrhoea and a confirmed case is a suspected case with *V. cholerae* O1 or O139 infection confirmed by culture.

Sample and metadata collection, and microbiological testing

Clinical samples, that is stool and rectal swabs, were collected in Yemen by epidemiological surveillance teams from suspected cholera cases during 2018 and 2019[11], and were transported to the NCPHL in the capital city Sana'a in Cary–Blair transport medium (Oxoid). To probe the diversity of vibrios shed by unreported cholera cases, as well as *V. cholerae* that may naturally occur in effluent waters, environmental samples were collected during the day time in October 2019 from the sewage system in and around Sana'a city and then transported to NCPHL for testing; each sample was collected in sterile bottles containing enrichment media comprised of 250 ml of sewage and alkaline peptone broth (Difco Laboratories) at a 1:1 ratio and incubated for 20 h at room temperature, including the transportation time to the NCPHL, and processed as described previously[26]. All samples were cultured and identified according to Centers for Disease Control and Prevention guidelines[27]. Resistance to antibiotics was tested by the disk diffusion method according to the Clinical & Laboratory Standards Institute guidelines[28] for a range of antibiotics as described in Supplementary Table 2.

Live clinical isolates ($n = 120$) were sent to IP, where only 21 samples were culture positive because of poor sample preservation during shipment (Supplementary Table 3 and Extended Data Fig. 3), leading to the final isolation of 22 *V. cholerae* strains (including two from mixed culture YE-NCPHL-18020). Strains re-isolated at IP were characterized by biochemical and serotyping methods according to standard practice of the French National Reference Centre for Vibrios and Cholera (CNRVC)[29]. Separate antibiotic susceptibility testing (Supplementary Tables 2 and 3) was performed by the disk diffusion method according to EUCAST guidelines[30] and minimum inhibitory concentration determination using the Sensititre (Thermo Fisher Scientific) and Etest (bioMérieux) systems. Interpretation into S (susceptible), I (intermediate) and R (resistant) categories was performed according to the 2020 edition of EUCAST recommendation on interpretation of the diameter of the zones of inhibition of *Enterobacteriaceae*[31], and to the 2013 Comité de l'Antibiogramme de la Société Française de Microbiologie standards for *Enterobacteriaceae*[32] for antibiotics for which critical diameters are no longer reported in the latest published guidelines. *E. coli* CIP 76.24 (ATCC 25922) was used as a reference strain.

DNA extraction and sequencing

Genomic DNA was extracted at the NCPHL from subcultures inoculated with single bacterial colonies and grown in nutrient agar (Oxoid) at 37 °C overnight according to the manufacturer's instructions (Wizard Genomic DNA Purification kit, Promega). Genomic DNA samples (derived from 10 environmental and 250 clinical samples, which includes the 120 samples sent to IP) were sent to the Wellcome Sanger Institute (WSI) and sequenced on the WSI sequencing pipeline (Extended Data Fig. 3) using the Illumina HiSeq platform X10 as described previously[33].

Genome assembly and annotation

The 260 sequencing read sets produced at the WSI (Extended Data Fig. 3) were processed with the WSI Pathogen Informatics pipeline[34]; quality of sequencing runs was assessed based on quality of mapping of 10% reads to the genome of reference strain N16961 (GenBank Assembly accession GCA_900205735.1) using the Burrows–Wheeler Aligner[35]; read sets passed the check if at least 80% bases were mapped after clipping, the base and indel error rate were smaller than 0.02, and fewer than 80% of the insert sizes fell within 25% of the most frequent size. Contamination was assessed manually based on Kraken classification of reads using the standard WSI Pathogen reference database, which contains all viral, archaeal and bacterial genomes and the mouse and human reference published in the RefSeq database as of 21 May 2015 (Supplementary Table 4). Sequences were assembled de novo into contigs as described previously[36], using SPAdes v.3.10.0 as the core assembler[37]. Poor assemblies were filtered out if differing by more than 20% from the expected genome size of 4.2 Mb, or when more than 10% of reads were assigned by Kraken to an organism other than *V. cholerae* (notably including the *Vibrio* phage ICP1) or to synthetic constructs, or were unclassified. This led to the exclusion of 28 genome assemblies, resulting in 232 high-quality assembled genomes. The genomes of strains CNRVC190243 and CNRVC190247 were assembled based on long and short reads using a hybrid approach with UniCycler[38] v.0.4.7 and v.0.4.8, respectively, using pilon[39] v.1.23 for the polishing step, to produce high-quality reference sequences comprised of both chromosomes and, for strain CNRVC190243, of an

additional plasmid, pCNRVC190243. New genomes were annotated with Prokka (v.1.5.0)[40].

Contextual genomic data

To provide phylogenetic context, we also included in this analysis previously published genome sequences from a globally representative set of isolates ('assembled *V. cholerae* genomes' dataset; $n = 882$). We first gathered genome assemblies generated at the WSI using the pipeline described above based on previously published short reads sets from *V. cholerae* isolates belonging to sub-lineage T13 of 7PET Wave 3 (7PET-T13) and from strains isolated in the close spatio-temporal context that is within a decade in Africa and South Asia (where the ancestor of T13 is thought to originate[7]). These include all 42 Yemen 2016–2017 isolates[7], 103 recent isolates from East Africa including from Kenya[7], Tanzania[41], Uganda[42] and Zimbabwe[16], and 74 isolates from South Asia[43]. In addition, we included genomes spanning the wider diversity of *V. cholerae*, including all 119 genomes from China[44], as well as 312 genomes from the collections of contextual genomes used in our previous studies[7,33]. Together with the 232 Yemen 2018–2019 isolate genome assemblies (see above), our final dataset consisted of 882 assembled *V. cholerae* genomes (Supplementary Table 5 and Extended Data Fig. 3).

Characterization of genomic features of interest

We first predicted the presence of various genomic features by searching genomes against reference databases. The presence of AMR genes, plasmid replicon regions or virulence factors were predicted using ABRicate[45], searching the reference databases NCBI AMR+[46], Plasmidfinder[47] or VFDB[48], respectively. We searched for the presence of conjugation apparatus (as a sign of plasmids or ICEs) and CRISPR–Cas arrays and subtyped these systems using MacSyFinder (v.2.1)[49] with models CONJScan (v.2.0.1) and CasFinder (v.3.1.0). Genomes positive for Cas systems were further analysed with CRISPRCasFinder[50] on the Pasteur Institute Galaxy server to retrieve CRISPR arrays.

In addition, we searched for sequences with significant similarity to previously described mobile elements relevant to our study. BLASTN[51] (v.2.7.1+, with default parameters) was used to identify known MGEs: the SXT ICE ICE*Vch*Ind5 (GenBank accession GQ463142.1); ICP1-like vibriophages ICP1_VMJ710 and ICP1_2012_A (GenBank accessions MN402506.2 and MH310936.1, respectively)[52] and the ICP1-like *Vibrio* phage YE-NCPHL-19021, which genome was the only assembled contig from the reads obtained from sample YE-NCPHL-19021 (this study; GenBank accession MW911613.1); the IncC-type plasmid pCNRVC190243, obtained from the hybrid assembly of strain CNRVC190243 described above (this study; ENA sequence accession OW443149.1); the MDR PCT Yem*Vch*MDRI, extracted from this plasmid (positions 16,442 to 36,862); PICI-like elements (PLE) 1, 2 and 3 (GenBank accessions KC152960.1, KC152961.1, MF176135.1)[53,54]. Absence of elements was verified at the read level as described below. Sequences similar to the reference sequences of the plasmid pCNRVC190243, the MDR PCT Yem*Vch*MDRI and the ICP1-like phage genome YE-NCPHL-19021 were also searched in a database of 661,405 genome assemblies[55] using a *k*-mer based COBS index[56]; alignment of best matches were further characterized using BLASTN.

In silico classification of *V. cholerae*

To predict the antigenic serogroup, we used BLASTN to screen the assemblies against a reference database of sequences of lipopolysaccharide O-antigen biosynthetic gene clusters that were delineated in the genome of reference strains for each known serogroup[57]. Best reference locus matches were identified as those with the highest combined score, summing scores of all local alignments, except when multiple local alignment overlapped in which case only best-scoring alignments were retained[58].

For O1 serotype prediction (Inaba or Ogawa), we used a combination of approaches including BLASTN search against the 882 assembled *V. cholerae* genomes (as described above) and ARIBA[59] v.2.14.6+ (with default parameters) to screen the sequencing read sets against the *wbeT* gene sequence from strain NCTC 9420 (positions 311,049–311,909 of GenBank accession CP013319.1) as a reference, as described previously[33]. Multilocus sequence typing of non-7PET isolates was conducted on PubMLST.org[60] under the non-O1/non-O139 *V. cholerae* seven-gene typing scheme.

Identification of single nucleotide variants

For variant calling, Illumina short reads from 7PET and *Vc*D genomes were mapped against the reference genomes from strains CNRVC190243 and CNRVC190247, respectively (456 'mapped 7PET genomes' and 33 'mapped *Vc*D genomes' datasets, respectively), and all genomes were mapped against the in-house MGE database described above. We mapped all 260 short-read sets from 2018–2019 Yemeni isolates sequenced at the WSI, including those 28 read sets that assembly showed low coverage or appeared contaminated with phage genomes (Supplementary Table 4); to recover variation data evidenced at the read level, provided reads were mapped at a sufficient depth (see below). We also mapped read sets from the 21 strains sequenced at the IP, and from contextual isolates of the 7PET-T13 sub-lineage and close relatives (Contextual genomic data), for a total of 468 mapped genomes. Reads were trimmed with Trimmomatic, mapped to both CNRVC190243 reference chromosomes with BWA-MEM and the IncC plasmid pCNRVC190243. Mapped genomes with an average read depth below 5× over the two chromosomes were deemed of insufficient read depth and were excluded (12 read sets mapped to CNRVC190243, all from this study and generated at WSI, were excluded for a final set of 456 mapped 7PET genomes (Supplementary Table 5); no read set mapped to CNRVC190247 was excluded). We used the software suite samtools/bcftools[61] v.1.9 to call single nucleotide variants with a minimum coverage of 10× read depth. Resulting consensus sequences were combined into a whole-genome alignment, which was processed with snp-sites[62] to produce an SNP alignment.

Overall genome similarity was assessed by computing SNP distances based on the above alignments using the function 'dist.dna' from the R package 'ape' [63], and ANI (accounting for unaligned regions) was computed using fastANI[64] v.1.3 with default parameters.

Phylogenetic inference

The Pantagruel pipeline[65] was used to infer a maximum-likelihood (ML) 'core-genome tree' using the '-S' option and otherwise default parameters. Briefly, 291 single-copy core-genome genes were extracted from the 882 assembled *V. cholerae* genomes; these 291 markers represent a strict definition of the core genome that is those genes occurring once and once only in all genomes of the dataset. This restrictive definition was intended to retain only genes with an expected high degree of sequence conservation and relatively low prevalence of horizontal gene transfer compared with other core genes, towards a robust phylogenetic inference at the species scale. The alignments of these 291 single-copy core-genome genes were concatenated and the resulting supermatrix was reduced to its 37,170 polymorphic positions, from which a ML tree was computed from RAxML[66] v.8.2.11 (model ASC_GTRGAMMAX using Stamatakis' ascertainment bias correction; one starting parsimony tree; 200 rapid bootstraps for estimating branch supports). Phylogenies were also inferred from whole-genome alignments of the concatenated consensus sequences of both chromosomes from the SNP alignment of the 456 mapped 7PET genomes and 33 mapped *Vc*D genomes. These alignments contained 2,092 and 91,312 polymorphic positions, respectively, and were used as input to RAxML-NG[67] v.1.0.1 to build the ML 'mapped genome trees' using the following options: 'all --tree pars{10} --bs-trees 200 --model GTR+G4+ASC_STAM'. Alternative

topologies were compared using RAxML-NG option '--sitelh' to generate per-site likelihood values and the 'SH.test' function from the 'phangorn' R package[68] to test hypotheses.

The 882 assembled *V. cholerae* core-genome tree was rooted using the clade of sequences identified as *V. paracholerae*[15] as an outgroup. The remaining part of the tree (*V. cholerae sensu stricto*) was subdivided into clades named *Vc*A to *Vc*K based on visual examination with the aim to coincide with previously described lineages such as 7PET and Gulf Coast among others, or based on even subdivisions of the tree diversity. *Vc*H, corresponding to the 7PET lineage, was further subdivided into clades of even depth, named subclades H.1 to H.9. The 456 mapped 7PET genomes were similarly classified into clusters based on the tree topology, with genomes assigned to subclades named *Vc*H.5, *Vc*H.6, *Vc*H.8 or *Vc*H.9 (according to their position in the 882 assembled *V. cholerae* core-genome tree). Genomes belonging to *Vc*H.9, which corresponds to the 7PET-T13 sub-lineage, were further separated into *Vc*H.9.a to *Vc*H.9.h, based on visual examination of the tree structure and aiming to maximize uniformity of the spatio-temporal metadata associated to genomes in each cluster; clusters correspond to clades, either entirely or at the exclusion of another cluster included in the clade; that is, genome clusters can emerge from each other. Final trees for the mapped genome datasets were rooted manually according to the branching pattern in the 882 assembled *V. cholerae* core-genome tree, the diversity of which encompasses that of the mapped genome trees.

From a subset of the 456 mapped 7PET genome alignments ($n = 335$) corresponding to *Vc*H.9, a recombination-free phylogeny (Extended Data Fig. 10) was inferred using ClonalFrameML[69] v.1.11 with default parameters and using the ML mapped genome tree (restricted to the *Vc*H.9 genome tips) as a starting tree. BactDating[70] v.1.1 was then used to estimate a timed phylogeny (using 100,000 Monte Carlo Markov chain iterations and otherwise default parameters) of the Yemen 2016–2019 genomes and relatives using the ClonalFrameML tree and day-resolved dates as input; median day of the year of isolation was used for isolates where these data were missing. Three independent chains were run from different random seeds and yielded close results.

Supporting data for phylogenetic analyses of the 882 assembled *V. cholerae*, 456 mapped 7PET genomes and 33 mapped *Vc*D genomes are available on the figshare repository (Data Availability).

Correlation of spatio-temporal and phylogenetic distances

GPS data associated with the site of sample collection (health centres) were used to compute spatial geodetic distances using R script 'gps_coords.r'[71,72]. Temporal distances were computed from the difference between day of collection (available only for 2018 and 2019 Yemen isolates). Phylogenetic distances were computed from the mapped genome tree using the function 'cophenetic' from the core R package 'stats'[73]. Spatial, temporal and phylogenetic distances were compared using a Monte Carlo approximation of the Mantel test as implemented in the 'mantel.randtest' function from the R package 'ade4'[74], using 100,000 permutations to compute the simulated *P* value. Maps showing the distribution of genomes clusters over the Yemen territory and in the region of Sana'a were obtained using QGIS 3.16.3 and the QuickOSM API to retrieve OpenStreetMap data, specifically level 4 administrative boundaries (governorates) in Yemen (last accessed 11 February 2021).

Pangenome analysis and clade-specific SNPs

On one hand, the synteny-aware pangenome pipeline Panaroo[75] v.1.2.3 was run on the 882 assembled *V. cholerae* genome set with the option '--clean-mode strict' and default parameters otherwise. On the other hand, a combined VCF file containing information on all SNP variation within the 456 mapped genome set was obtained using the 'bcftools merge' command. To identify clade-specific SNPs and accessory gene presence/absence patterns, we used custom R scripts[58] to compare the combined VCF file and the gene presence/absence table output of Panaroo, respectively, with the mapped genome tree. Based on lists of genomes assigned to various clades and clusters (Results), we identify SNPs or accessory genes that are specific of a focus clade in contrast to a background group or a sister clade, considering the contrast significant when the Bonferroni-corrected *P* value is below 0.05 and when the frequency of an allele is above 0.8 in the focus clade and below 0.2 in the background clade, or conversely. Pangenome analysis files are available on figshare (https://doi.org/10.6084/m9.figshare.19519105). Putative anti-phage defence systems were searched by testing correlation of presence/absence patterns between ICP1-like phage and each pangenome gene cluster; only associations with Pearson correlation coefficients lower than −0.9 or greater than 0.9 and *P* values lower than 10^{-5} were retained as significant.

Ethics and approval of sampling

This study is based exclusively on bacterial isolates and derived genomic DNA extracts and complies with all relevant ethical regulations as follows. None of the human samples from which the strains were isolated were collected specifically for this study, as all were collected as part of the cholera outbreak surveillance effort led by the NCPHL. The metadata associated with the samples and retained in the study are age and sex, as well as place of hospitalization, which do not identify the patient and do not warrant informed consent or ethics committee approval. Bacterial isolate cultures were later sent to IP, while genomic DNA extracts were transferred to the WSI. We note that Yemen is not a signatory to the Nagoya Protocol and therefore does not require transfer authorization under international law.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Short-read genomic data sequenced at the WSI were deposited at the European Nucleotide Archive (ENA) under the BioProject PRJEB34436. Four of the resulting assemblies comprised a single 123-kb contig corresponding to the ICP1-like phage; these assemblies were deemed uncontaminated and complete ICP1-like phage genomes and were deposited to GenBank under the accessions MW911612–MW911615. Complete hybrid genome assemblies for reference strains CNRVC019243 and CNRVC019247 were deposited to the ENA under the BioProject accessions PRJEB52123 and PRJEB47951 (Assemblies GCA_937000105 and GCA_937000115), respectively. Supplementary data are available online on the Figshare repository, under the following digital object identifiers (doi): https://doi.org/10.6084/m9.figshare.16595999, https://doi.org/10.6084/m9.figshare.16611823, https://doi.org/10.6084/m9.figshare.18304961, https://doi.org/10.6084/m9.figshare.19097111, https://doi.org/10.6084/m9.figshare.19519105, https://doi.org/10.6084/m9.figshare.23653971, https://doi.org/10.6084/m9.figshare.23849034.

Code availability

All custom code used in this study was made available in a git repository publicly availably on GitHub at https://github.com/flass/yemenpaper (release v.0.3; https://doi.org/10.5281/zenodo.8221344).

References

1. *UNHCR Yemen: 2021 Country Operational Plan* (UNHCR, 2021); https://data2.unhcr.org/en/documents/details/85850
2. Dureab, F. et al. Assessment of electronic disease early warning system for improved disease surveillance and outbreak response in Yemen. *BMC Public Health* **20**, 1422 (2020).
3. *Cholera* (WHO EMRO, 2021); http://www.emro.who.int/health-topics/cholera-outbreak/cholera-outbreaks.html

4. Health workers in Yemen reach more than 306,000 people with cholera vaccines during four-day pause in fighting—WHO, UNICEF. *WHO* https://www.who.int/news/item/05-10-2018-health-workers-in-yemen-reach-more-than-306-000-people-with-cholera-vaccines-during-four-day-pause-in-fighting-who-unicef (2018).
5. Federspiel, F. & Ali, M. The cholera outbreak in Yemen: lessons learned and way forward. *BMC Public Health* **18**, 1338 (2018).
6. Mutreja, A. et al. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
7. Weill, F.-X. et al. Genomic insights into the 2016–2017 cholera epidemic in Yemen. *Nature* **565**, 230–233 (2019).
8. Bai, Z. G. et al. Azithromycin versus penicillin G benzathine for early syphilis. *Cochrane Database Syst. Rev.* https://doi.org/10.1002/14651858.CD007270.pub2 (2012)
9. Rabaan, A. A. Cholera: an overview with reference to the Yemen epidemic. *Front. Med.* **13**, 213–228 (2019).
10. *Technical Note on the Use of Antibiotics for the Treatment and Control of Cholera* (Global Task Force on Cholera Control, 2018); https://www.gtfcc.org/wp-content/uploads/2019/10/gtfcc-technical-note-on-use-of-antibiotics-for-the-treatment-of-cholera.pdf
11. Camacho, A. et al. Cholera epidemic in Yemen, 2016–18: an analysis of surveillance data. *Lancet Glob. Health* **6**, e680–e690 (2018).
12. Ambrose, S. J., Harmer, C. J. & Hall, R. M. Compatibility and entry exclusion of IncA and IncC plasmids revisited: IncA and IncC plasmids are compatible. *Plasmid* **96–97**, 7–12 (2018).
13. De, R. Mobile genetic elements of *Vibrio cholerae* and the evolution of its antimicrobial resistance. *Front. Trop. Dis.* **2**, 7 (2021).
14. Lassalle, F. & Bashir, I. M. Epidemiological line lists NCPHL – *V. cholerae* identification and antibiotic susceptibility testing from suspected cholera patient samples. *figshare* https://doi.org/10.6084/m9.figshare.23653971.v1 (2023).
15. Islam, M. T. et al. Population analysis of *Vibrio cholerae* in aquatic reservoirs reveals a novel sister species (*Vibrio paracholerae* sp. nov.) with a history of association with human infections. *Appl. Environ. Microbiol.* **87**, e0042221 (2021).
16. Mashe, T. et al. Highly resistant cholera outbreak strain in Zimbabwe. *N. Engl. J. Med.* **383**, 687–689 (2020).
17. Spagnoletti, M. et al. Acquisition and evolution of SXT-R391 integrative conjugative elements in the seventh-pandemic *Vibrio cholerae* lineage. *mBio* **5**, e01356-14 (2014).
18. Harmer, C. J., Pong, C. H. & Hall, R. M. Structures bounded by directly-oriented members of the IS26 family are pseudo-compound transposons. *Plasmid* **111**, 102530 (2020).
19. Bonnin, R. A. et al. PER-7, an extended-spectrum β-lactamase with increased activity toward broad-spectrum cephalosporins in *Acinetobacter baumannii*. *Antimicrob. Agents Chemother.* **55**, 2424–2427 (2011).
20. Toleman, M. A., Bennett, P. M. & Walsh, T. R. ISCR elements: novel gene-capturing systems of the 21st century? *Microbiol. Mol. Biol. Rev.* **70**, 296–316 (2006).
21. Yu, L. et al. Multiple antibiotic resistance of *Vibrio cholerae* serogroup O139 in China from 1993 to 2009. *PLoS ONE* **7**, e38633 (2012).
22. Wang, R. et al. IncA/C plasmids harboured in serious multidrug-resistant *Vibrio cholerae* serogroup O139 strains in China. *Int. J. Antimicrob. Agents* **45**, 249–254 (2015).
23. Opazo, A. et al. Plasmid-encoded PER-7 β-lactamase responsible for ceftazidime resistance in *Acinetobacter baumannii* isolated in the United Arab Emirates. *J. Antimicrob. Chemother.* **67**, 1619–1622 (2012).
24. Ceccarelli, D., Salvia, A. M., Sami, J., Cappuccinelli, P. & Colombo, M. M. New cluster of plasmid-located class 1 integrons in *Vibrio cholerae* O1 and a *dfrA15* cassette-containing integron in *Vibrio parahaemolyticus* isolated in Angola. *Antimicrob. Agents Chemother.* **50**, 2493–2499 (2006).
25. Jaskólska, M., Adams, D. W. & Blokesch, M. Two defence systems eliminate plasmids from seventh pandemic *Vibrio cholerae*. *Nature* **604**, 323–329 (2022).
26. Madico, G. et al. Active surveillance for *Vibrio cholerae* O1 and vibriophages in sewage water as a potential tool to predict cholera outbreaks. *J. Clin. Microbiol.* **34**, 2968–2972 (1996).
27. *Laboratory Methods for the Diagnosis of Vibrio cholerae* (Centers for Disease Control and Prevention, 2021); https://www.cdc.gov/cholera/pdf/laboratory-methods-for-the-diagnosis-of-vibrio-cholerae-chapter-4.pdf
28. Weinstein, M. P. *M100: Performance Standards for Antimicrobial Susceptibility Testing* 31st edn (Clinical & Laboratory Standards Institute, 2021).
29. Dodin, A. & Fournier, J. M. in *Diagnosis of the Cholera Vibrio* 59–82 (Institut Pasteur, 1992).
30. *Recommandations 2020* version 1.1 (CA-SFM & EUCAST, 2020); https://www.sfm-microbiologie.org/wp-content/uploads/2020/04/CASFM2020_Avril2020_V1.1.pdf
31. *Breakpoint Tables for Interpretation of MICs and Zone Diameters* version 10.0 (EUCAST, 2020); https://www.eucast.org/fileadmin/src/media/PDFs/EUCAST_files/Breakpoint_tables/v_10.0_Breakpoint_Tables.pdf
32. *Recommandations 2013* (CA-SFM, 2013); https://www.sfm-microbiologie.org/wp-content/uploads/2020/07/CASFM_2013.pdf
33. Dorman, M. J. et al. Genomics of the Argentinian cholera epidemic elucidate the contrasting dynamics of epidemic and endemic *Vibrio cholerae*. *Nat. Commun.* **11**, 4918 (2020).
34. WSI Pathogen Informatics. vr-codebase. *GitHub* (2022); https://github.com/sanger-pathogens/vr-codebase
35. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997 (2013).
36. Page, A. J. et al. Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb. Genom.* **2**, e000083 (2016).
37. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
38. Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput. Biol.* **13**, e1005595 (2017).
39. Walker, B. J. et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS ONE* **9**, e112963 (2014).
40. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
41. Kachwamba, Y. et al. Genetic characterization of *Vibrio cholerae* O1 isolates from outbreaks between 2011 and 2015 in Tanzania. *BMC Infect. Dis.* **17**, 157 (2017).
42. Bwire, G. et al. Molecular characterization of *Vibrio cholerae* responsible for cholera epidemics in Uganda by PCR, MLVA and WGS. *PLoS Negl. Trop. Dis.* **12**, e0006492 (2018).
43. Morita, D. et al. Whole-genome analysis of clinical *Vibrio cholerae* O1 in Kolkata, India, and Dhaka, Bangladesh, reveals two lineages of circulating strains, indicating variation in genomic attributes. *mBio* **11**, e01227-20 (2020).
44. Wang, H. et al. Genomic epidemiology of *Vibrio cholerae* reveals the regional and global spread of two epidemic non-toxigenic lineages. *PLoS Negl. Trop. Dis.* **14**, e0008046 (2020).

45. Seemann, T. ABRicate. *GitHub* (2021); https://github.com/tseemann/abricate
46. Feldgarden, M. et al. Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype–phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**, e00483-19 (2019).
47. Carattoli, A. et al. In silico detection and typing of plasmids using plasmidfinder and plasmid multilocus sequence typing. *Antimicrob. Agents Chemother.* **58**, 3895–3903 (2014).
48. Chen, L., Zheng, D., Liu, B., Yang, J. & Jin, Q. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. *Nucleic Acids Res.* **44**, D694–D697 (2016).
49. Néron, B. et al. MacSyFinder v2: Improved modelling and search engine to identify molecular systems in genomes. *Peer Community J.* **3**, e28 (2023).
50. Pourcel, C. et al. CRISPRCasdb a successor of CRISPRdb containing CRISPR arrays and cas genes from complete genome sequences, and tools to download and query lists of repeats and spacers. *Nucleic Acids Res.* **48**, D535–D544 (2020).
51. Camacho, C. et al. BLAST+: architecture and applications. *BMC Bioinform.* **10**, 421 (2009).
52. Angermeyer, A., Das, M. M., Singh, D. V. & Seed, K. D. Analysis of 19 highly conserved *Vibrio cholerae* bacteriophages isolated from environmental and patient sources over a twelve-year period. *Viruses* **10**, 299 (2018).
53. Seed, K. D., Lazinski, D. W., Calderwood, S. B. & Camilli, A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. *Nature* **494**, 489–491 (2013).
54. O'Hara, B. J., Barth, Z. K., McKitterick, A. C. & Seed, K. D. A highly specific phage defense system is a conserved feature of the *Vibrio cholerae* mobilome. *PLoS Genet.* **13**, e1006838 (2017).
55. Blackwell, G. A. et al. Exploring bacterial diversity via a curated and searchable snapshot of archived DNA sequences. *PLoS Biol.* **19**, e3001421 (2021).
56. Bingmann, T., Bradley, P., Gauger, F. & Iqbal, Z. COBS: a compact bit-sliced signature index. In *String Processing and Information Retrieval. SPIRE 2019* (eds Brisaboa, N. & Puglisi, S.) 285–303 (Springer, 2019).
57. Murase, K. et al. Genomic dissection of the *Vibrio cholerae* O-serogroup global reference strains: reassessing our view of diversity and plasticity between two chromosomes. *Microb. Genom.* **8**, mgen000860 (2022).
58. Lassalle, F. flass/yemenpaper. *GitHub* (2022); https://github.com/flass/yemenpaper
59. Hunt, M. et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb. Genom.* **3**, e000131 (2017).
60. Jolley, K. A., Bray, J. E. & Maiden, M. C. J. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res.* **3**, 124 (2018).
61. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
62. Page, A. J. et al. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genom.* **2**, e000056 (2016).
63. Paradis, E. *Analysis of Phylogenetics and Evolution with R* (Springer, 2012).
64. Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat. Commun.* **9**, 5114 (2018).
65. Lassalle, F., Jauneikaite, E., Veber, P. & Didelot, X. Automated reconstruction of all gene histories in large bacterial pangenome datasets and search for co-evolved gene modules with Pantagruel. Preprint at *bioRxiv* https://doi.org/10.1101/586495 (2019).
66. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
67. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).
68. Schliep, K. P. phangorn: phylogenetic analysis in R. *Bioinformatics* **27**, 592–593 (2011).
69. Didelot, X. & Wilson, D. J. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. *PLoS Comput. Biol.* **11**, e1004041 (2015).
70. Didelot, X., Croucher, N. J., Bentley, S. D., Harris, S. R. & Wilson, D. J. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res.* **46**, e134 (2018).
71. Lassalle, F. flass/microbiomes. *GitHub* (2018); https://github.com/flass/microbiomes
72. Lassalle, F. et al. Oral microbiomes from hunter-gatherers and traditional farmers reveal shifts in commensal balance and pathogen load linked to diet. *Mol. Ecol.* **27**, 182–195 (2018).
73. R Core Team *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2020).
74. Dufour, A.-B. & Dray, S. The ade4 Package: implementing the duality diagram for ecologists. *J. Stat. Softw.* **22**, 1–20 (2007).
75. Tonkin-Hill, G. et al. Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol.* **21**, 180 (2020).

Acknowledgements

This research was funded in whole, or in part, by the Wellcome Trust (grant nos 206194 and 108413/A/15/D). This work was supported by Institut Pasteur, Santé publique France, and by the French Government's Investissement d'Avenir programme, Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (grant no. ANR-10-LABX-62-IBEID). F.L., M.J.D., G.A.B., A.T.-B., M.A.B., A.C. and N.R.T. were supported by Wellcome funding to the Sanger Institute (grant nos 206194 and 108413/A/15/D). E.N., J.R., M.-L.Q. and F.-X.W. were supported by French Government funding to Institute Pasteur (grant no. ANR-10-LABX-62-IBEID). We thank J. Woolfolk, S. Clare and C. Tolley for their support in sample management at Wellcome Sanger Institute, as well as the Sanger Pipelines team for support. M.J.D. is an Official Fellow of Churchill College, Cambridge, and was previously supported by a Junior Research Fellowship at the College. For the purpose of Open Access, the author has applied a CC-BY public copyright licence to any author accepted manuscript version arising from this submission.

Author contributions

F.L., A.A.-H., M.-L.Q., F.-X.W., G.D. and N.R.T. conceived the study. S.A.-S., M.A.H., E.N., I.M.B., M.J.D., A.A.A.-S., A.A.-M., K.A., M.A., A.A.-H., M.-L.Q. and G.D. were involved in the acquisition and/or preparation of samples. F.L., M.J.D., J.R., G.A.B., A.T.-B., M.A.B., A.C., M.-L.Q. and F.-X.W. analysed and/or interpreted the data. F.L. created new software used in the work. F.L., M.J.D., A.T.-B., M.A.B., A.C., M.-L.Q., F.-X.W., G.D. and N.R.T. drafted the work and/or substantially revised it. All authors agree to the submitted version and approved substantial modifications as appropriate and agree to be accountable for their contributions.

Competing interests

The authors declare no competing interests.

Additional information

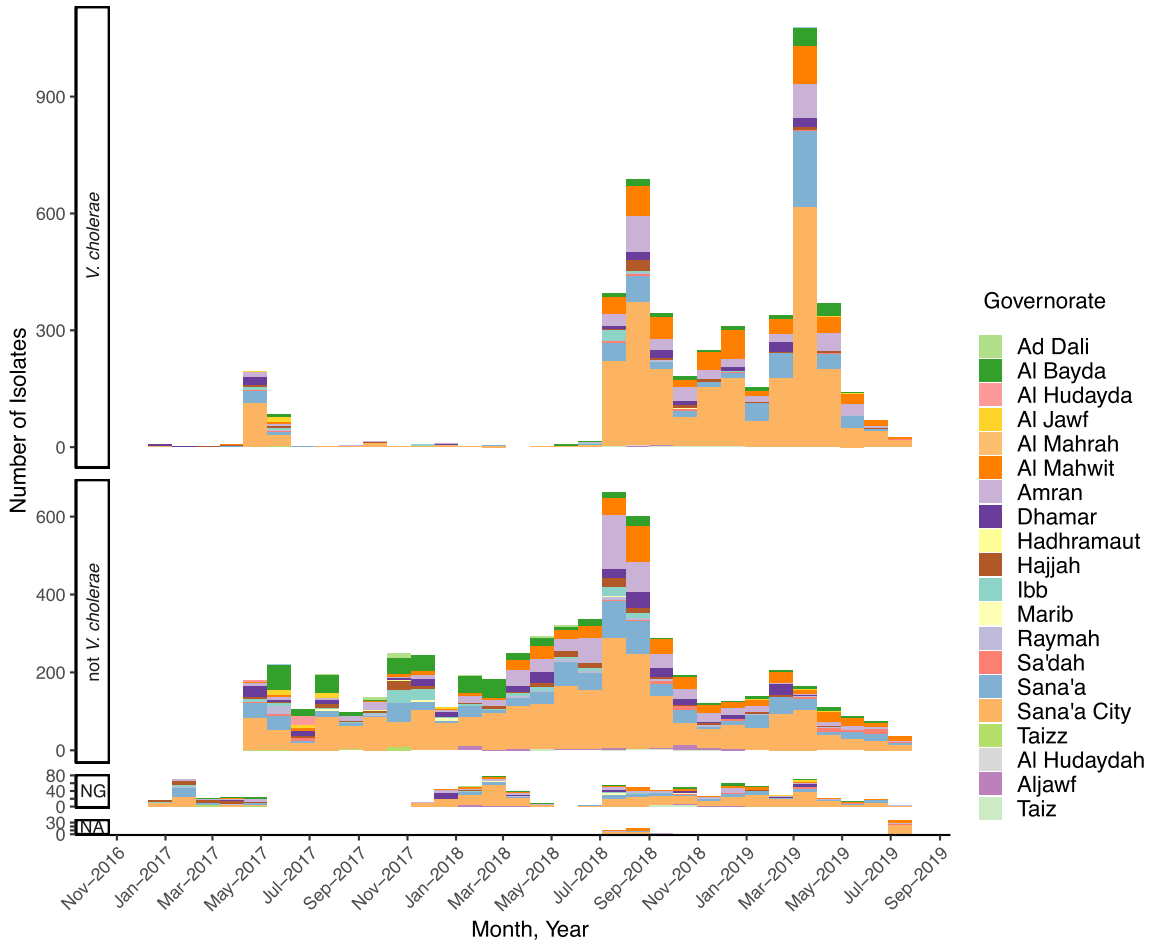
Extended data is available for this paper at https://doi.org/10.1038/s41564-023-01472-1.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41564-023-01472-1.

Correspondence and requests for materials should be addressed to Florent Lassalle, Ghulam Dhabaan or Nicholas R. Thomson.

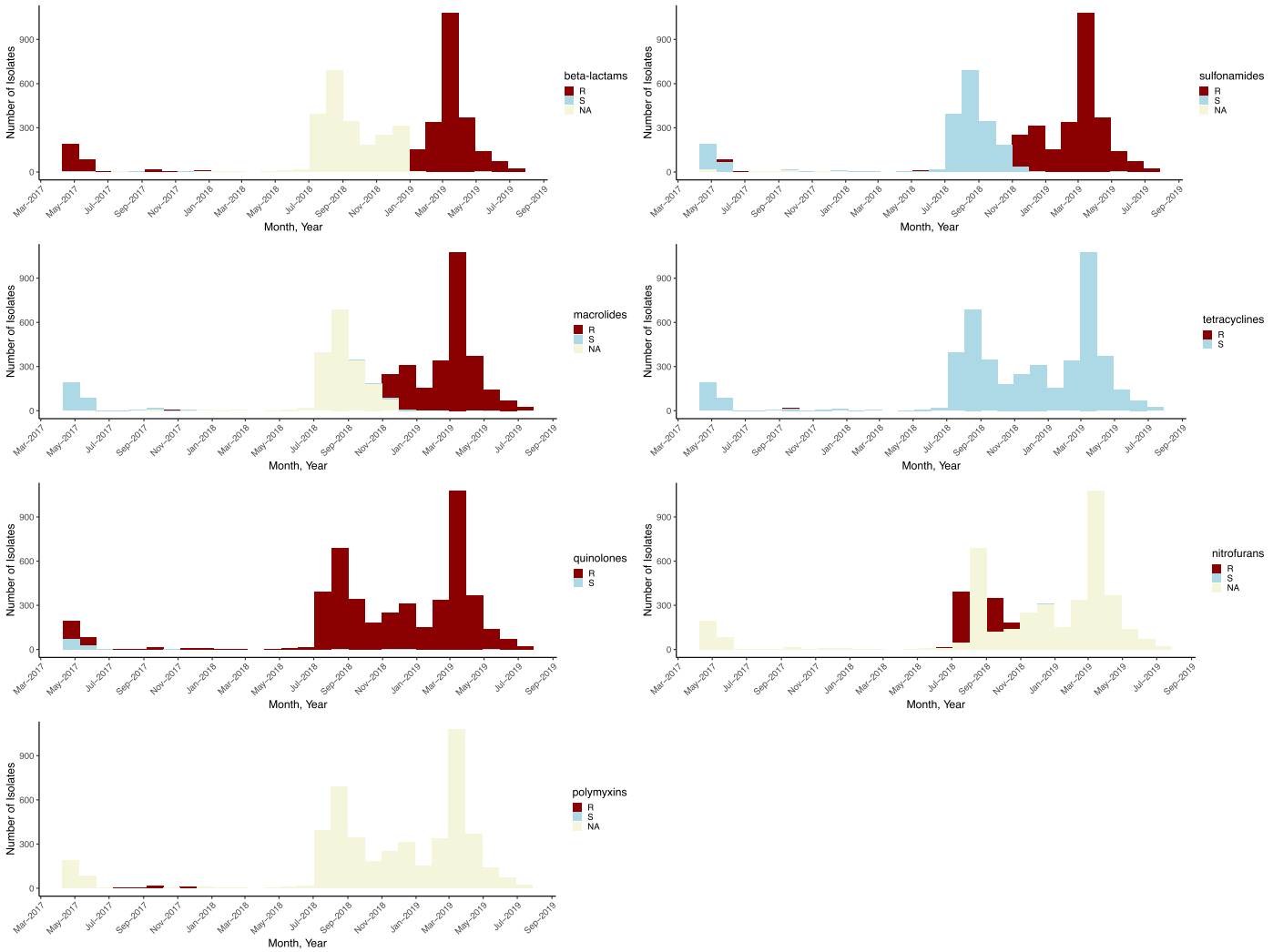
Peer review information *Nature Microbiology* thanks Pedro Oliveira and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

https://doi.org/10.1038/s41564-023-01472-1



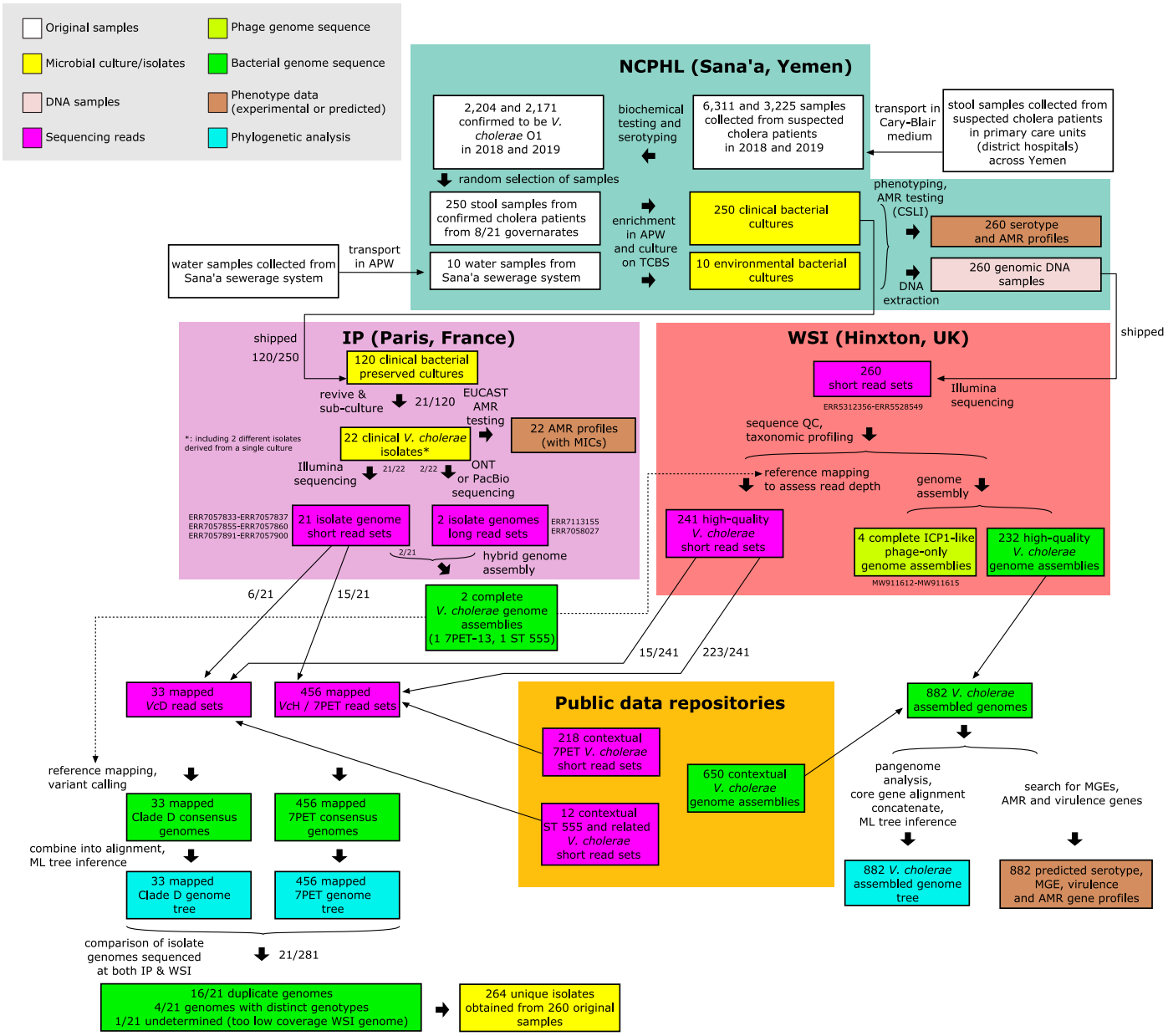
Extended Data Fig. 1 | Culture confirmation of samples derived from suspected cholera cases in Yemen, 2017–2019. Distribution over time of *V. cholerae* culture result samples received at the NCPHL, broken down by governorate. Data are derived from Electronic Disease Early Warning System (eDEWS) sample lists (Supplementary Table 1). NG, no growth; NA, not available.

https://doi.org/10.1038/s41564-023-01472-1



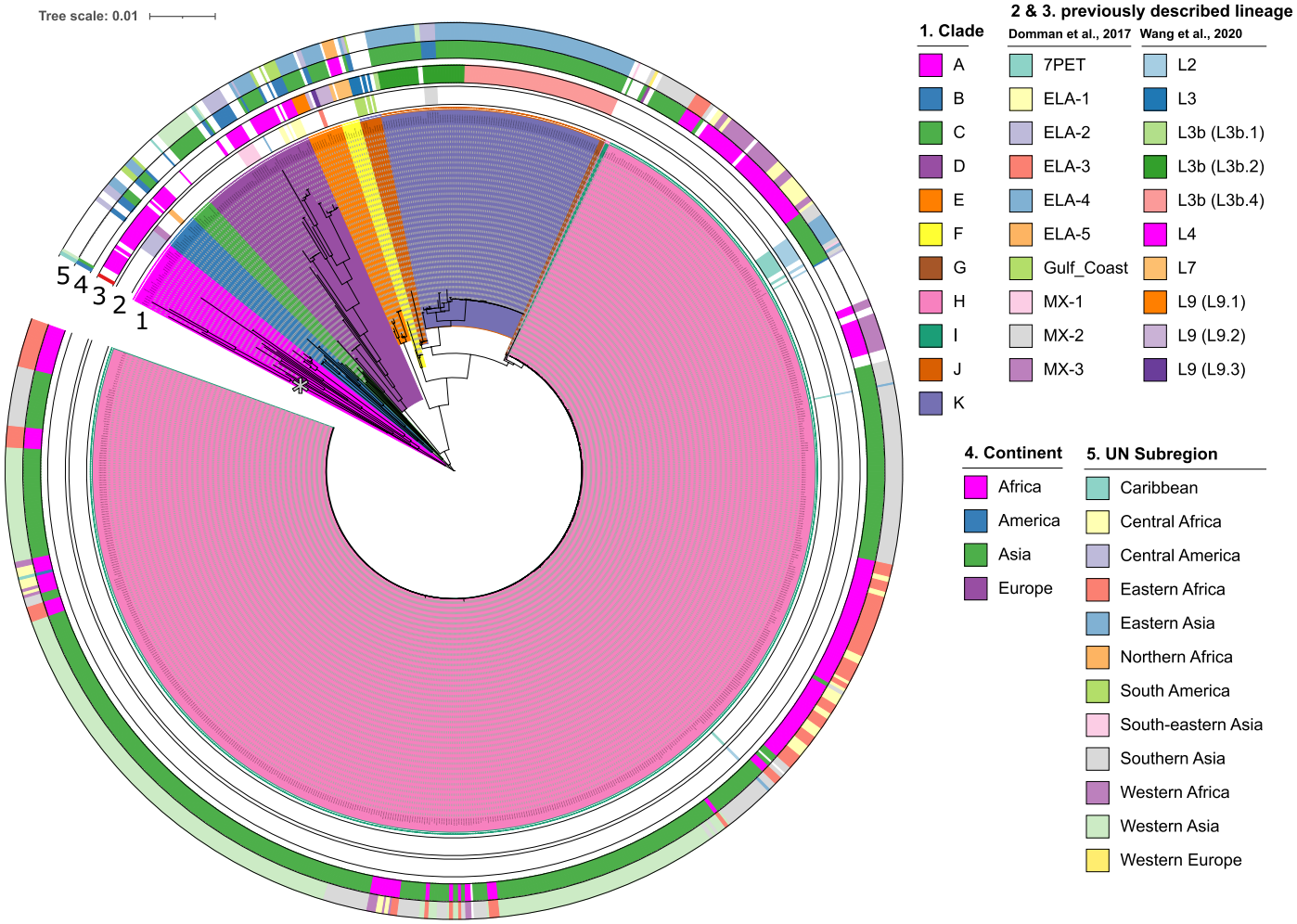
Extended Data Fig. 2 | Antibiotic susceptibility phenotypes of all *V. cholerae* isolates collected in Yemen, 2017 and 2019. Distribution over time of resistance and sensitivity to broad antibiotic classes among culture-confirmed *V. cholerae* isolates received at the NCPHL. Data are derived from eDEWS sample lists (Supplementary Table 1).

https://doi.org/10.1038/s41564-023-01472-1



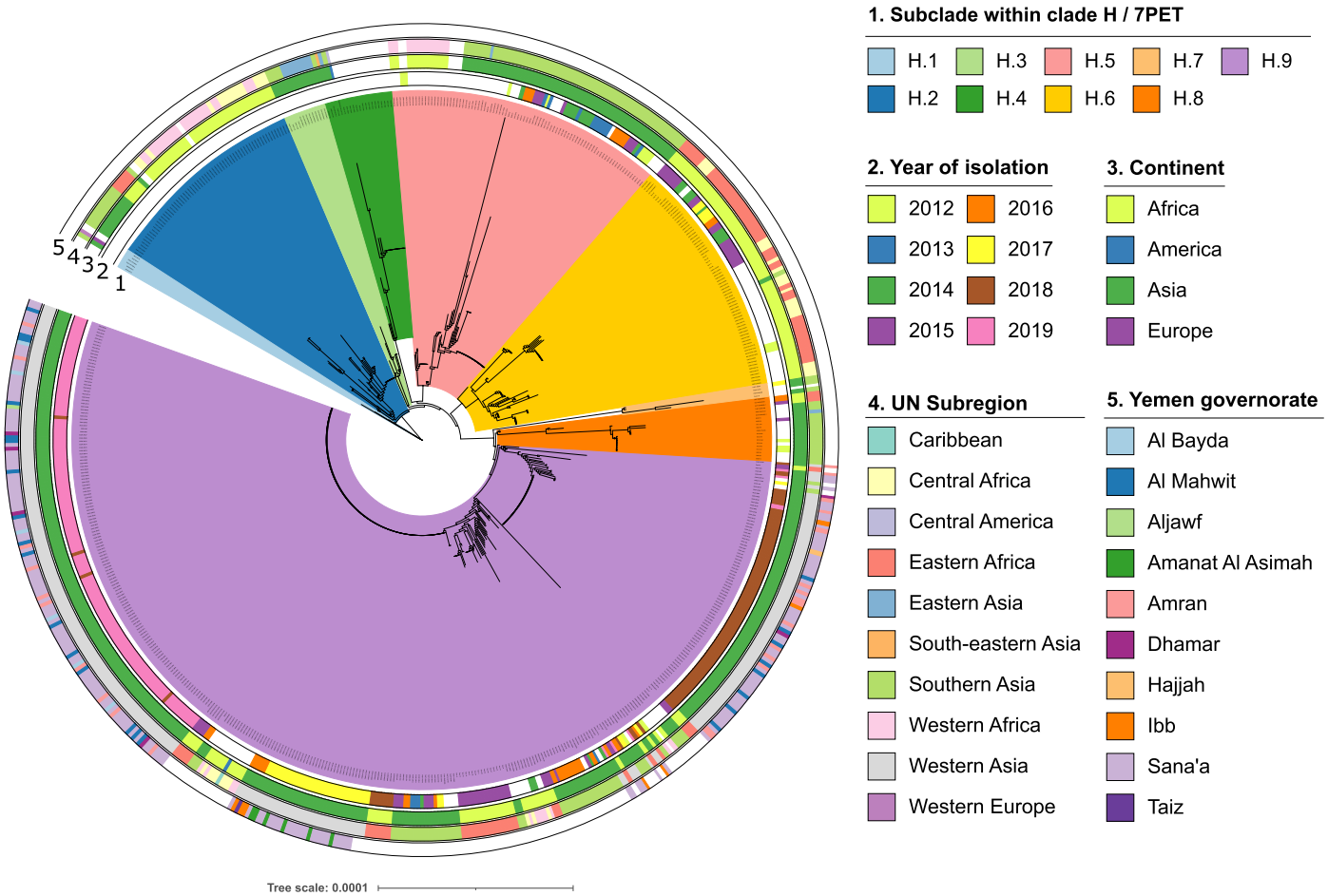
Extended Data Fig. 3 | Flowchart of sample collection, management and use in protocols and analyses. Experiments and analyses are itemized and grouped according to the different locations of the collaborative consortium where they were undertaken: NCPHL, The National Centre of Public Health Laboratories; IP, Institut Pasteur; WSI, Wellcome Sanger Institute.

https://doi.org/10.1038/s41564-023-01472-1



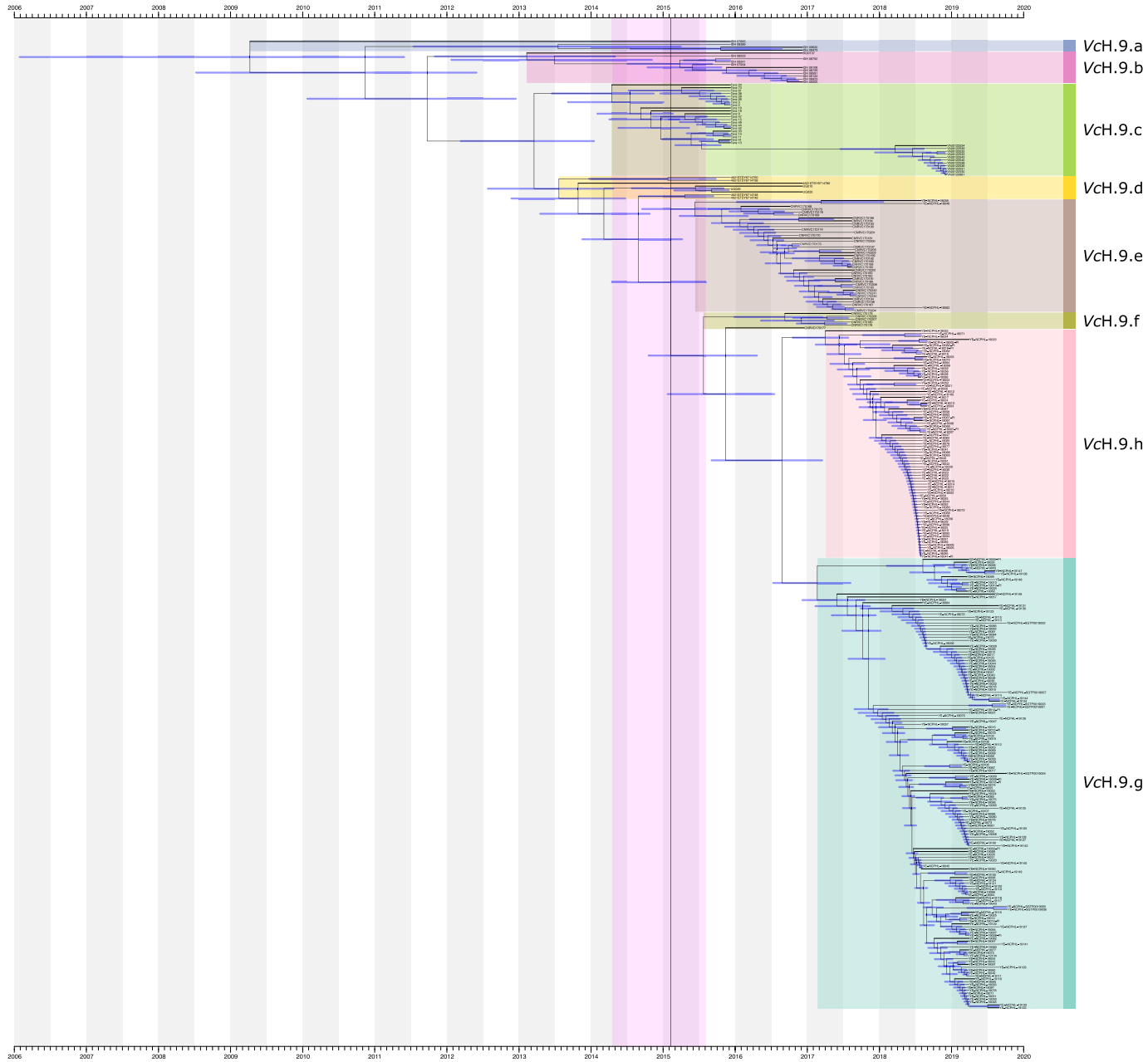
Extended Data Fig. 4 | Phylogenetic diversity of *Vibrio cholerae* isolates from Yemen and contextual samples. Expanded version of phylogentic tree shown in Fig. 1 (no clades are collapsed).

https://doi.org/10.1038/s41564-023-01472-1



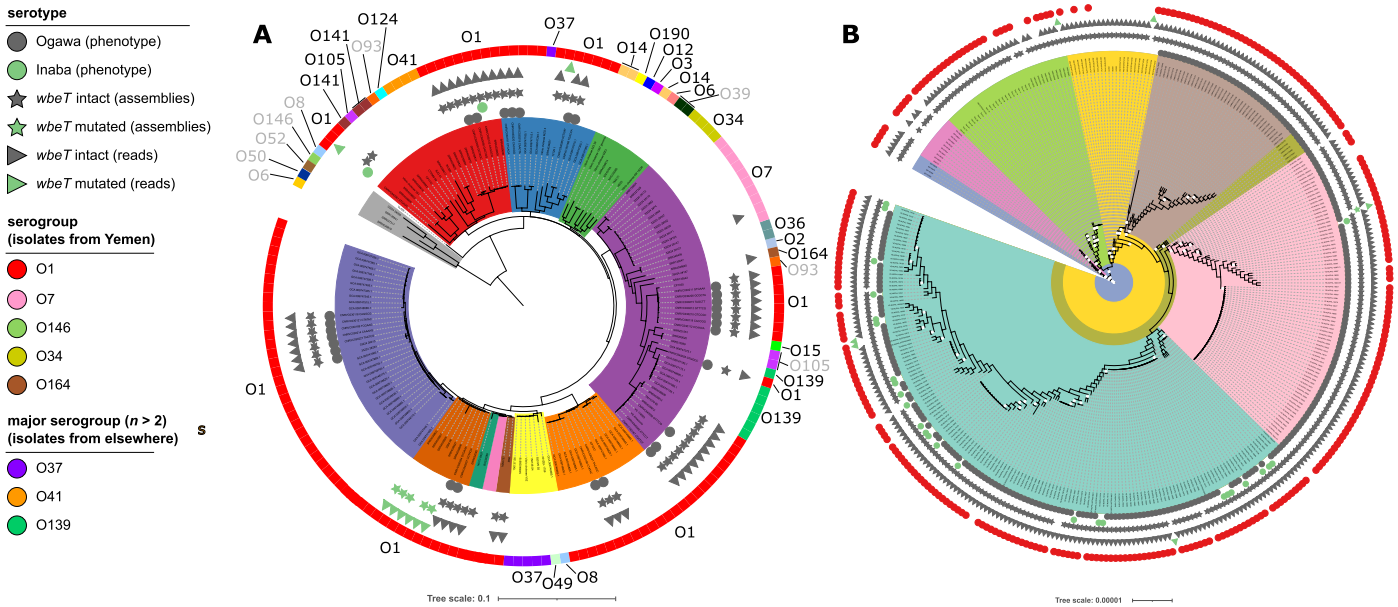
Extended Data Fig. 5 | Phylogenetic diversity of *Vibrio cholerae Vc*H isolates from Yemen and contextual samples. Expanded version of phylogentic tree shown in Fig. 1, focusing on the subtree of clade *Vc*H (collapsed in Fig. 1), with details of its phylogenetic substructure.

https://doi.org/10.1038/s41564-023-01472-1



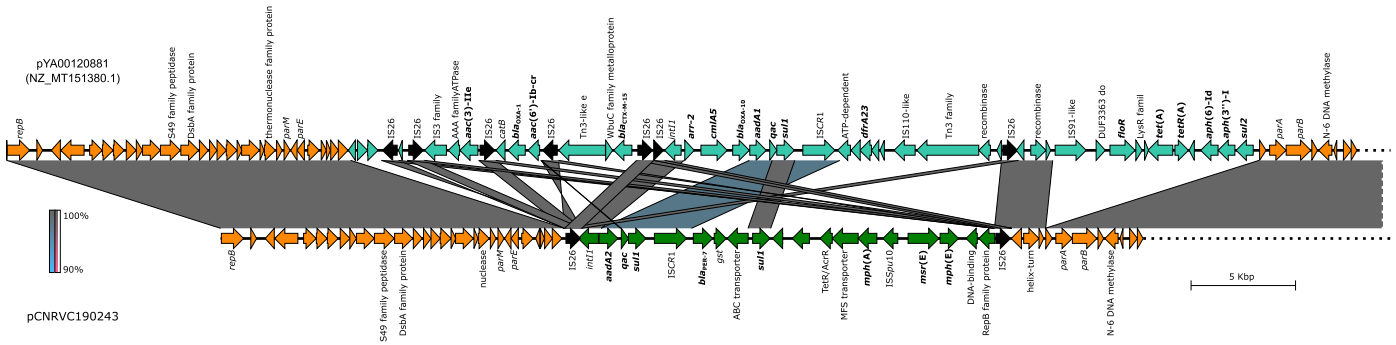
Extended Data Fig. 6 | Timed phylogeny of *Vibrio cholerae Vc*H.9 isolates from Yemen and contextual samples. Timed phylogeny of 335 *Vc*H.9 genomes obtained by estimating the dates of nodes using BactDating from a recombination-free phylogeny computed with ClonalFrameML as input; the 335 genomes correspond to the *Vc*H.9 clade within the 456 mapped 7PET genome tree (as presented in Fig. 2a). Subclusters are labelled and coloured as per previous figures. X axis represents time in years. Horizontal blue bars indicate 95% confidence intervals around the node dates. The vertical purple line marks the emergence of the clade of Yemen isolates.

https://doi.org/10.1038/s41564-023-01472-1



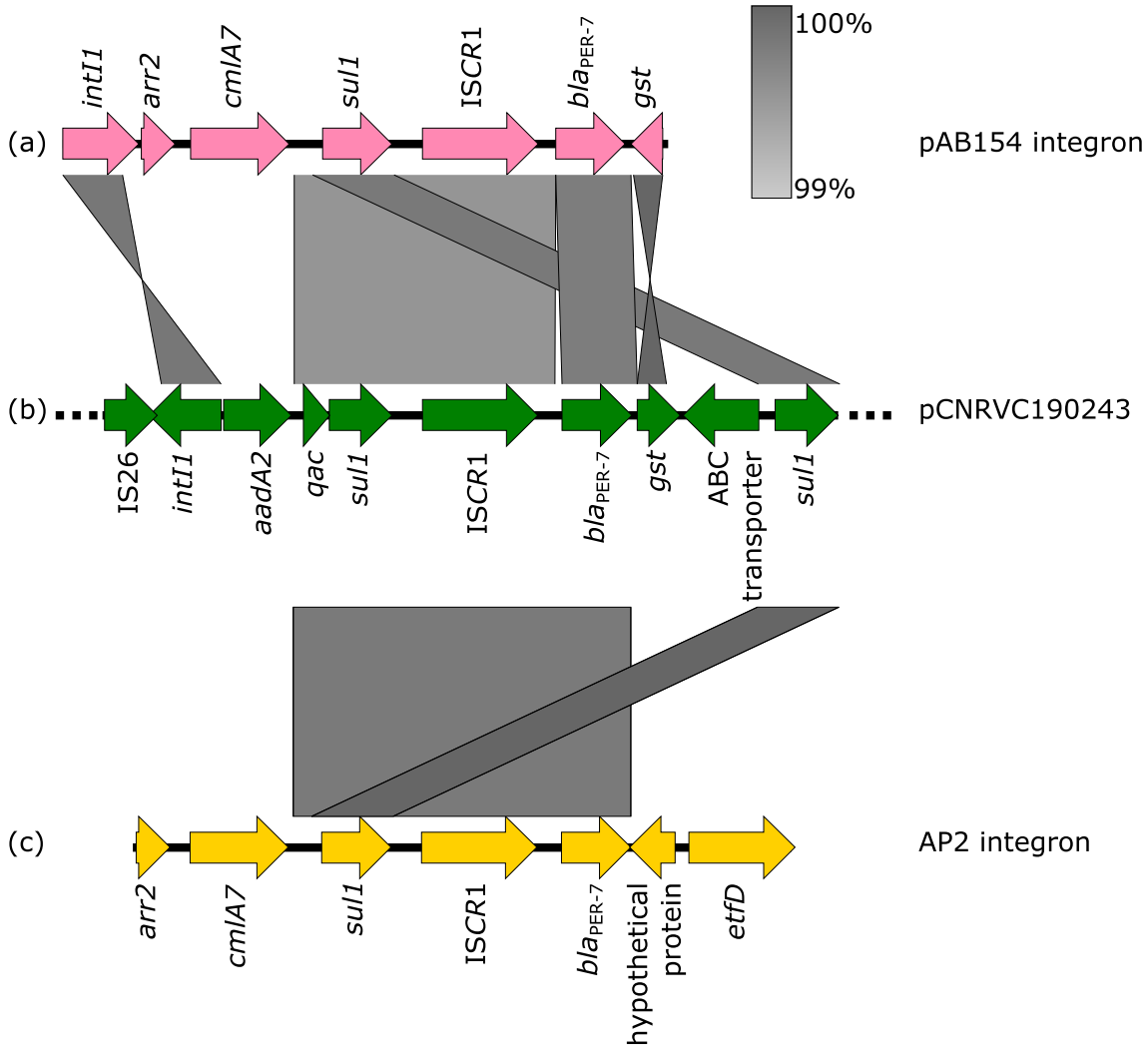
Extended Data Fig. 7 | *In-silico* prediction of the O-antigen diversity among *Vibrio cholerae* from Yemen and contextual samples. Prediction of lipopolysaccharide (LPS) O-antigen serogroup and O1 serotypes projected on the isolate trees corresponding to **(A)** the core-genome tree as presented in Fig. 1a,b the subtree of the mapped 7PET genome tree as presented in Fig. 2a. Serogroup prediction are based on a best normalized BlastN score in a search against the reference LPS O-antigen biosynthetic cluster sequence from Murase et al. 2022; predictions where the best hit had a normalized percent nucleotide identity below 98% are indicated in grey font.

https://doi.org/10.1038/s41564-023-01472-1



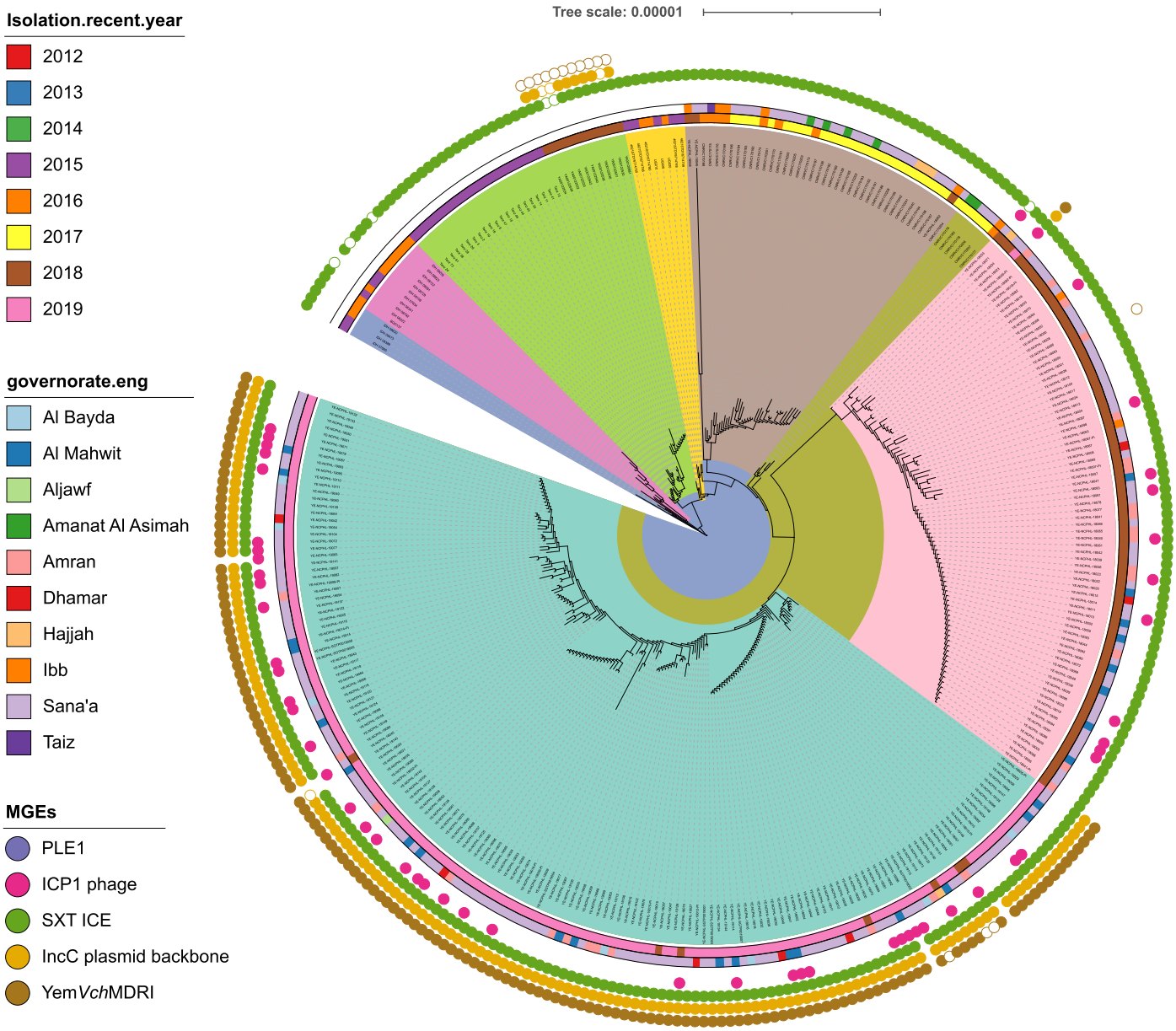
Extended Data Fig. 8 | Comparison of IncC plasmids pCNRVC190243 and pYA00120881. BlastN alignment of the differing regions of IncC plasmids pCNRVC190243 and pYA00120881. Regions of similarity are highlighted by interleaving bands; band colour indicate similarity intensity and orientation of the alignment (blue, direct; pink, reverse).

https://doi.org/10.1038/s41564-023-01472-1



Extended Data Fig. 9 | Comparison of IS*CR*1 elements. BlastN alignments of IS*CR*1 regions of (**a**) *A. baumanii* str. AB154 plasmid pAB154 integron (JQ639792.1), (**b**) pCNRVC190243/Yem*Vch*MDRI and (**c**) *A. baumanii* str. AP2 integron (HQ713678.1).

https://doi.org/10.1038/s41564-023-01472-1



Extended Data Fig. 10 | Recombination-free phylogeny of *Vibrio cholerae* VcH.9 isolates from Yemen and contextual samples. Tree computed from the same alignment as in Fig. 2a, but using ClonalFrameML to infer a recombination-free phylogeny reflecting the clonal propagation of the organism.

nature portfolio

Corresponding author(s): Florent Lassalle, Ghulam Dhabaan, Nick Thomson

Last updated by author(s): Mar 3, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	no software was used for data collection
Data analysis	genome assembly: SPAdes v3.10.0, UniCycler v0.4.7 and v0.4.8, pilon v1.23; genome annotation: Prokka version v1.5.0; sequence similarity searches: NCBI BLAST+ v2.7.1, Abricate v1.0.1, MacSyFinder v2.1, CRISPRCasFinder v1.1.2, ARIBA v2.14.6+, samtools/bcftools v1.9; phylogenetic analysis: RAxML-NG v1.0.1, Pantagruel version 8f95544, ClonalFrameML v1.11, BactDating v1.1; pangenome analysis: Panaroo82 v1.2.3; statistical analysis: R with packages 'ade4' and 'stats', custom code in https://github.com/flass/yemenpaper; geographical maps: QGIS 3.16.3.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

Short-read genomic data sequenced at the WSI were deposited at the ENA under the BioProject PRJEB34436. Four of the resulting assemblies comprised a single 123-kb contig corresponding to the ICP1-like phage; these assemblies were deemed uncontaminated and complete ICP1-like phage genomes and were deposited to GenBank under the accessions MW911612-MW911615. Complete hybrid genome assemblies for reference strains CNRVC019243 and CNRVC019247 were deposited to the ENA under the BioProject accessions PRJEB52123 and PRJEB47951 (Assemblies GCA_937000105 and GCA_937000115), respectively. Supplementary data are available online on the Figshare repository, under the following digital object ientifiers (doi): https://doi.org/10.6084/m9.figshare.16595999, https://doi.org/10.6084/m9.figshare.16611823, https://doi.org/10.6084/m9.figshare.18304961, https://doi.org/10.6084/m9.figshare.19097111, https://doi.org/10.6084/m9.figshare.19519105, https://doi.org/10.6084/m9.figshare.23653971, https//:doi.org/10.6084/m9.figshare.23849034.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	this section is not relevant as the study did not focus on the human aspect of cholera patients but on the bacterial pathogen; no human material was collected or studied.
Population characteristics	this section is not relevant as the study did not focus on the human aspect of cholera patients but on the bacterial pathogen; no human material was collected or studied.
Recruitment	this section is not relevant as the study did not focus on the human aspect of cholera patients but on the bacterial pathogen; no human material was collected or studied.
Ethics oversight	No ethics were required as the samples were not human material

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Among the clinical samples collected from suspected cholera patients in Yemen in from 2016 to 2019, presence of Vibrio cholerae was tested by microbiological culture and upon positive identification antibiotic susceptibility was performed on isolates. The changes in the antibiotic susceptibility pattern between 2018 and 2019 of these isolates prompted us to randomly choose 260 V. cholerae isolates from both years for whole genome sequencing (WGS) towards a genomic epidemiology analysis.
Research sample	260 V. cholerae isolates were chosen for WGS. 250 isolates were derived from clinical samples chosen randomly among the 4,375 samples confirmed to be positive for V. cholerae O1 by culture in 2018 and 2019 in Yemen. 10 additional isolates were derived from environmental samples obtained from the sewer system in Sana'a in 2019.
Sampling strategy	No calculation were done to establish adequate sample size. Sample size was determined due on the limited resources at the NCPHL lab in Sana'a in the context of the ongoing war and humanitarian crisis.
Data collection	Metadata related to the clinical samples were collected through the Electronic Disease Early Warning System (eDEWS), a surveillance programme coordinated by the Ministry of Public Health and Population of Yemen (MPHP) in Sana'a used to monitor the epidemic.
Timing and spatial scale	Samples sent for WGS were chosen randomly among a collection of samples obtained throughout the outbreak in 2018 and 2019, with spatio-temporal density of the sample roughly reflecting the variations in intensity of the outbreak through time and space. These samples originated from eight of the 21 Yemen governorates, comprising 71 out of 333 districts (Table S1), with 101 samples collected in 2018 (from mid-July to late October) and 149 in 2019 (from late February to late April and from early August to mid-October). In addition, ten environmentally-derived strains were isolated from sewerage in Sana'a in October 2019.

Data exclusions	Poor genome assemblies were filtered out if differing of more than 20% from the expected genome size of 4.2 Mb, or when more than 10% of reads were assigned by Kraken to another organism than V. cholerae (notably including the Vibrio phage ICP1) or to synthetic constructs, or were unclassified. This led to the omission of 28 genome assemblies, resulting in 232 high-quality assembled genomes to be included in the 882 assembled V. cholerae genomes dataset. Mapped genomes with an average read depth below 5x over the two chromosomes were deemed of insufficient read depth and were excluded (12 read sets mapped to CNRVC190243, all from this study and generated at WSI, were excluded for a final set of 456 mapped 7PET genomes; no read set mapped to CNRVC190247 was excluded).
Reproducibility	20 samples were sequenced twice, once at the Wellcome Sanger Institute (WSI; Hinxton, UK) and once at the Institut Pasteur (IP; Paris, France). For the isolates derived from these 20 samples, antibiotic susceptibility testing (AST) was also done twice, once at the National Centre of Public Health Laboratories (NCPHL; Sana'a, Yemen), once at IP. Discrepancies of outcome occurred, as 4 genomes derived from the same original sample were of different genotype; we explained this by the presence of multiple V. cholerae strains within the samples, an hypothesis confirmed by PCR testing of the samples. These isolates with distinct genotypes obtained from the same samples were then treated as separate isolates in downstream analyses.
Randomization	This is not relevant to our study, as groups of bacterial isolates were determined based on their genotype using phylogenetic analysis.
Blinding	The phylogenetic trees were initially drawn without any geographic information associated with the genomes.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging