LETTER

https://doi.org/10.1038/s41586-018-0818-3

Genomic insights into the 2016–2017 cholera epidemic in Yemen

François-Xavier Weill[1,21]*, Daryl Domman[2,3,21], Elisabeth Njamkepo[1], Abdullrahman A. Almesbahi[4], Mona Naji[4], Samar Saeed Nasher[4], Ankur Rakesh[5], Abdullah M. Assiri[6], Naresh Chand Sharma[7], Samuel Kariuki[8], Mohammad Reza Pourshafie[9], Jean Rauzier[1], Abdinasir Abubakar[10], Jane Y. Carter[11], Joseph F. Wamala[12], Caroline Seguin[13], Christiane Bouchier[14], Thérèse Malliavin[15], Bita Bakhshi[16], Hayder H. N. Abulmaali[17], Dhirendra Kumar[7,18], Samuel M. Njoroge[8], Mamunur Rahman Malik[10], John Kiiru[8], Francisco J. Luquero[5], Andrew S. Azman[19], Thandavarayan Ramamurthy[18], Nicholas R. Thomson[2,20,22] & Marie-Laure Quilici[1,22]

Yemen is currently experiencing, to our knowledge, the largest cholera epidemic in recent history. The first cases were declared in September 2016, and over 1.1 million cases and 2,300 deaths have since been reported[1]. Here we investigate the phylogenetic relationships, pathogenesis and determinants of antimicrobial resistance by sequencing the genomes of *Vibrio cholerae* isolates from the epidemic in Yemen and recent isolates from neighbouring regions. These 116 genomic sequences were placed within the phylogenetic context of a global collection of 1,087 isolates of the seventh pandemic *V. cholerae* serogroups O1 and O139 biotype El Tor[2–4]. We show that the isolates from Yemen that were collected during the two epidemiological waves of the epidemic[1]—the first between 28 September 2016 and 23 April 2017 (25,839 suspected cases) and the second beginning on 24 April 2017 (more than 1 million suspected cases)—are *V. cholerae* serotype Ogawa isolates from a single sublineage of the seventh pandemic *V. cholerae* O1 El Tor (7PET) lineage. Using genomic approaches, we link the epidemic in Yemen to global radiations of pandemic *V. cholerae* and show that this sublineage originated from South Asia and that it caused outbreaks in East Africa before appearing in Yemen. Furthermore, we show that the isolates from Yemen are susceptible to several antibiotics that are commonly used to treat cholera and to polymyxin B, resistance to which is used as a marker of the El Tor biotype.

We investigated the bacterial populations that caused the cholera epidemic in Yemen by sequencing 42 *V. cholerae* O1 serotype Ogawa isolates that were recovered during this epidemic. Thirty-nine of these isolates were collected from patients with cholera who lived in three different governorates of Yemen (Fig. 1a, b). They span both waves of the epidemic, having been collected between 5 October 2016 and 31 August 2017. The three remaining isolates were collected from patients from a temporary refugee centre on the Saudi Arabia–Yemen border on 30 August 2017 (Fig. 1b). We also sequenced 74 7PET isolates from South Asia, the Middle East, and Eastern and Central Africa (Extended Data Fig. 1 and Supplementary Table 1). We placed these new isolates in the context of a global collection of 1,087 7PET genomic sequences[2–4] (Supplementary Table 1) and constructed a maximum-likelihood phylogeny of 1,203 genomes, using 9,986 single-nucleotide variants (SNVs) that were evenly distributed across the non-repetitive, non-recombinant core genome (Fig. 2a).

We also detected a strong temporal signal, making it possible to estimate time-scaled phylogenies (Fig. 2b and Extended Data Figs. 2–4), which showed that the epidemic in Yemen originated from a recently emerged 7PET wave 3 clade[5], which contains the cholera toxin subunit B gene variant *ctxB7* (Fig. 2). All of the isolates from Yemen clustered together (median pairwise SNV difference of 3 (range, 0–13)), confirming that the two epidemiological waves that were observed during the epidemic in Yemen, which had very different clinical attack rates[1], were produced by a single clone rather than arising from two separate introductions. We estimated the date of the most recent common ancestor of the isolates from Yemen to January 2016 (95% Bayesian credible interval, September 2015 to June 2016) (Fig. 2b, Extended Data Fig. 3 and Extended Data Table 1). Our phylogenetic analysis shows that the isolates from Yemen are different from those that have been circulating in the Middle East over the last decade, such as those isolated in Iraq in 2007 and 2015, and in Iran from 2012 to 2015 (Fig. 2a). These isolates from the Middle East also belong to 7PET wave 3, but are attributed to different sublineages on the phylogenetic tree and were imported from South Asia on two separate occasions. The isolates from Yemen are most closely related to isolates that were collected from outbreaks in Eastern Africa (Kenya, Tanzania[3] and Uganda[4]) from 2015 to 2016 (Fig. 2). Collectively, these isolates belong to a new sublineage (T13), which corresponds to the most recent, newly identified introduction of 7PET into East Africa. All of these T13 isolates are different from those previously recovered in West or East Africa (sublineages T12 and T10, respectively) (Fig. 2). Our data suggest that this 7PET wave 3 clade, which contains all isolates with the *ctxB7* allele, first emerged in South Asia in the early 2000s (Fig. 2b), consistent with the first detection of *ctxB7* isolates in Kolkata, India in 2006[6]. This *ctxB7* clade has been exported to areas outside Asia on at least three separate occasions: West Africa (T12 introduction event)[2] in 2008 (estimates with 95% credible interval), Haiti in 2010[7,8] and East Africa (T13 introduction event)[4] between 2013 and 2014 (estimates with 95% credible interval) (Fig. 2b, Extended Data Fig. 3 and Extended Data Table 1).

In addition to the *ctxB7* allele, all of the analysed isolates from Yemen had the following genomic features (Table 1): (1) the toxin-coregulated pilus gene subunit A gene variant $tcpA^{CIRS101}$; (2) a deletion (ΔVC0495–VC0512) within *Vibrio* seventh pandemic island II (VSP-II); and (3) an SXT/R391-integrating conjugating element (ICE) called ICE*Vch*Ind5/ICE*Vch*Ban5, which is associated with multiple-drug resistance.

[1]Institut Pasteur, Unité des Bactéries Pathogènes Entériques, Paris, France. [2]Wellcome Sanger Institute, Hinxton, UK. [3]Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA. [4]National Centre of Public Health Laboratories (NCPHL), Sana'a, Yemen. [5]Epicentre, Paris, France. [6]Ministry of Health, Riyadh, Saudi Arabia. [7]Maharishi Valmiki Infectious Diseases Hospital, Delhi, India. [8]Centre for Microbiology Research, Kenya Medical Research Institute, Nairobi, Kenya. [9]Pasteur Institute of Iran, Department of Bacteriology, Tehran, Iran. [10]WHO Regional Office for the Eastern Mediterranean (EMRO), Cairo, Egypt. [11]Amref Health Africa, Nairobi, Kenya. [12]WHO, Juba, South Sudan. [13]Médecins Sans Frontières (MSF), Dubai, United Arab Emirates. [14]Institut Pasteur, Plate-forme Génomique (PF1), Paris, France. [15]Unité de Bioinformatique Structurale, UMR 3528, CNRS; C3BI, USR 3756, Institut Pasteur, Paris, France. [16]Department of Bacteriology, Faculty of Medical Sciences, Tarbiat Modares University, Tehran, Iran. [17]Central Public Health Laboratory (CPHL), Baghdad, Iraq. [18]Translational Health Science and Technology Institute (THSTI), Faridabad, Haryana, India. [19]Department of Epidemiology, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA. [20]London School of Hygiene and Tropical Medicine, London, UK. [21]These authors contributed equally: François-Xavier Weill, Daryl Domman. [22]These authors jointly supervised this work: Nicholas R. Thomson, Marie-Laure Quilici.
*e-mail: francois-xavier.weill@pasteur.fr

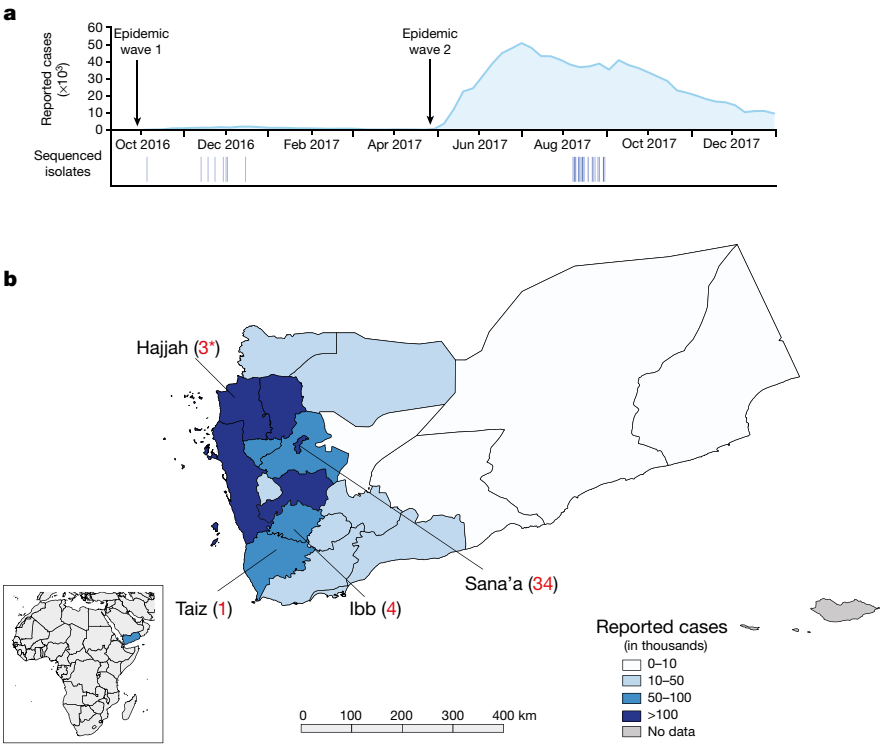


Fig. 1 | Geographical location of the sequenced *V. cholerae* O1 El Tor isolates and number of reported cholera cases. a, Aggregate number of suspected cholera cases per week in Yemen until 31 December 2017 (http://yemeneoc.org/bi/), showing the two epidemic waves. The dates of the isolates sequenced in this study are shown under the epidemic curve. **b**, Geographical location of the 42 *V. cholerae* O1 El Tor isolates from Yemen. The three isolates collected in Saudi Arabia (denoted by the asterisk) were obtained from Yemeni refugees from Hajjah District and are considered to be 'Yemeni isolates' throughout the manuscript. The number of cases per governorate is indicated according to a previous study[1]. The governate map of Yemen was created using QGIS version 2.16 (https://qgis.org) and the shape file was approved for use by the UN Office for the Coordination of Humanitarian Affairs (OCHA), OCHA Yemen country office (https://data.humdata.org/dataset/yemen-admin-boundaries). The small inlay map was created using QGIS version 2.16 using the Natural Earth base map version 4.0.0 (https://www.naturalearthdata.com).

Consistent with the genomic evidence, all of the isolates from Yemen have a similar narrow phenotype of antimicrobial drug resistance to nalidixic acid, the vibriostatic agent O/129 and nitrofurantoin (Table 1). Mutations in the DNA gyrase gene *gyrA* that resulted in an S83I amino acid substitution and mutations in the topoisomerase IV gene *parC* that resulted in an S85L substitution explain the resistance of the isolates from Yemen to nalidixic acid and their decreased susceptibility to ciprofloxacin. An approximately 10-kb deletion in ICE variable region III resulted in the loss of four genes that encode resistance to streptomycin (*strA* and *strB*), chloramphenicol (*floR)* and sulfonamides (*sul2)*. The fifth gene of this region, which encodes resistance to the vibriostatic agent O/129 (*dfrA1)*, is present in the isolates from Yemen. This deletion is not unique, as similar deletions that encompass the *strA*, *strB*, *floR* and *sul2* genes, flanked by transposase genes, have independently arisen several times in 7PET wave 3 isolates[2,9]. The resistance of *V. cholerae* to nitrofurans is due to the loss of expression of a reductase enzyme that converts the drug into its active form[10]. By combining phenotypic and genotypic data, we found lesions in the *VC0715* and *VCA0637* genes of nitrofuran-resistant isolates (Extended Data Table 2). *VC0715* and *VCA0637* encode orthologues of the NfsA (52% amino acid identity) and NfsB (58% amino acid identity) proteins of *Escherichia coli* K12 (GenBank accession number NC_000913), respectively. In *E. coli*, disruption of the nitroreductases that are encoded by these genes confers nitrofuran resistance[11]. In all 7PET wave 3 isolates, including the isolates from Yemen, the observed mutations in *VC0715* led to a R169C amino acid substitution and the mutation in *VCA0637* introduced a premature stop codon (Q5Stop) that probably abolishes protein function.

The isolates from Yemen were also susceptible to polymyxins. This is an important finding, because resistance to polymyxin B has been used as a marker of the *V. cholerae* O1 El Tor biotype since the beginning of the seventh cholera pandemic in 1961[12,13]. Unlike the El Tor biotype, the classical biotype (responsible for the six previous pandemics)[14] is susceptible to polymyxin B. Polymyxin resistance is conferred by changes to the lipid A domain of the surface lipopolysaccharide, thereby altering its charge[12,13]. The *vprA* (*VC1320*) gene, disruption of which is known to restore susceptibility to polymyxin in 7PET isolates, is required for expression of the *almEFG* operon that encodes the genes that are required for the glycine modification of lipid A[12]. A specific non-synonymous SNV in *vprA* genes (predicted to result in a D89N substitution, Extended Data Fig. 5) was present in 97% (63 out of 65) of polymyxin B-susceptible isolates (Extended Data Table 2), including all of the isolates from Yemen. The first polymyxin-susceptible 7PET isolates with this VprA D89N substitution in our dataset were identified in South Asia in 2012 (Fig. 2b), consistent with microbiological data from Kolkata, India, where polymyxin B-susceptible *V. cholerae* O1 isolates emerged in 2012 and replaced polymyxin-resistant strains after 2014[15].

7PET isolates from the *ctxB7* clade have been associated with the two largest cholera epidemics in recent history. In addition to the current Yemeni epidemic, the introduction of this sublineage into Haiti in 2010 in the wake of a devastating earthquake, resulted in one million cases and almost 10,000 deaths by 2017[16,17]. These two major events highlight the threat that cholera continues to pose to public health in vulnerable populations. The UN (United Nations) estimates that 16 out of 29 million people in Yemen lack access to clean water and basic sanitation because of the destruction of public and health infrastructures during the years of civil conflict[18]. The complexity of the situation in Yemen before the epidemic was set against a backdrop of large acute watery diarrhoea or cholera outbreaks across the Horn of Africa (Extended Data Fig. 1), which serves as a major hub of migration into Yemen[19,20]. This region, which links Asia to Africa at the southern entrance of the Red Sea, has long been a crossroads of trade and communication routes. Several importations of 7PET cholera from Asia into the Horn of Africa are likely to have followed this route, such as T3 in 1970[2].

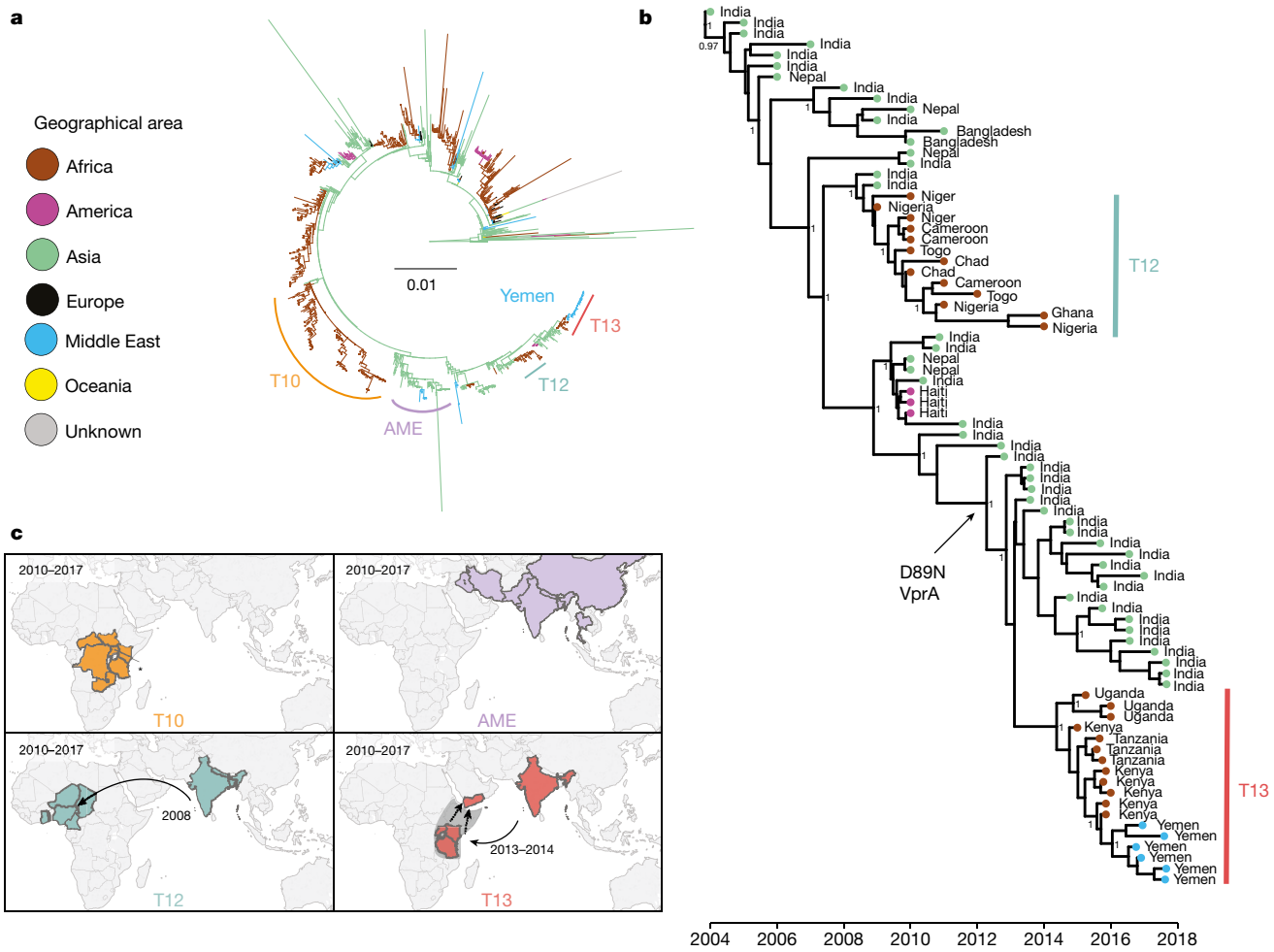


Fig. 2 | Phylogenetic relatedness of the *V. cholerae* O1 El Tor isolates from the 2016–2017 epidemic in Yemen. a, Maximum-likelihood phylogeny of 1,203 genomic sequences. M66 was used as the outgroup. The scale bar denotes substitutions per variable site (SNVs). Branches are coloured according to geographical location, inferred by stochastic mapping of the geographical origin of each isolate onto the tree. The inferred introduction events into Africa are indicated by the letter 'T'. The sublineage labelled AME (Asia/Middle East) contains the most recent Middle Eastern isolates. **b**, Maximum clade credibility tree produced with BEAST for a subset of 81 representative isolates of the distal part of genomic wave 3 (that is, those with the *ctxB7* allele). Geographical location of the isolates is indicated in the same colours as in **a**. Selected nodes supported by posterior probability values ≥0.8 are shown. Acquisition of the polymyxin susceptibility-associated non-synonymous SNV in VC1320 (*vprA*) is indicated. **c**, The geographical distribution of selected 7PET sublineages. An asterisk denotes data from a previous study[4]. The date ranges shown for introductions are the 95% credible interval estimate of the most recent common ancestor in years. Dashed lines and the grey area in T13 indicate that the sublineage was detected in East Africa before its appearance in Yemen. This does not represent a precise route of transmission. The maps were created with Tableau Desktop version 10.1.5 using the base map from © OpenStreetMap contributors (https://www.openstreetmap.org), available under an Open Database License.

Table 1 | Characteristics of the 2016–2017 cholera epidemic strain from Yemen

Species	*Vibrio cholerae*
Serogroup, serotype and biotype	O1, Ogawa, El Tor
Genomic wave	3
Genetic markers	*ctxB7*, *tcpA*^{CIRS101}, *rtxA*°, VSP-II$^{\Delta}$
AMR profile	POL^{S} (1–2 mg l^{-1}), COL^{S} (2–8 mg l^{-1}), O129^{R}, NAL^{R} (64 to ≥256 mg l^{-1}), CIP^{DS} (0.25–0.5 mg l^{-1}), FT^{R}
Horizontally acquired AMR element, acquired AMR gene (AMR phenotype)	ICE*Vch*Ind5/ICE*Vch*Ban5$^{\Delta}$, *dfrA1* (O129^{R})
Mutated chromosomal genes (AMR phenotype)	*gyrA* S83I and *parC* S85L (NAL^{R}, CIP^{DS}) *nfsA* R169C and *nfsB* Q5Stop (FT^{R})

MIC range values are indicated in parentheses. The superscript R, S and DS indicate resistant, susceptible and decreased susceptibility, respectively. VSP-II$^{\Delta}$ indicates a deletion that encompasses VC0495–VC0512, according to GenBank accession AE003852. ICE*Vch*Ind5/ICE*Vch*Ban5$^{\Delta}$ indicates a deletion that encompasses ICE*Vch*Ind50011–ICE*Vch*InD50021 according to GenBank accession GQ463142. O129^{R} denotes cross-resistance to the vibriostatic agent O/129 and trimethoprim. AMR, antimicrobial resistance; CIP, ciprofloxacin; COL, polymyxin E; FT, nitrofurantoin; NAL, nalidixic acid; POL, polymyxin B.

The available genomic data for the historical and current importations of the 7PET sublineage into Africa are not consistent with a local origin, but instead highlight the importance of human-mediated spread of the epidemic 7PET lineage from South Asia. An inability to obtain samples from countries in this region hampered our efforts to reconstruct the routes of transmission in East Africa before the appearance of this strain in Yemen more precisely.

In summary, a single recent 7PET sublineage with an unusual antimicrobial resistance phenotype is responsible for the cholera epidemic in Yemen. Our study illustrates the key role of genomic microbial surveillance and cross-border collaborations in understanding the global spread of cholera, the evolution of virulence and determinants of antibiotic resistance.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, statements of data availability and associated accession codes are available at https://doi.org/10.1038/s41586-018-0818-3.

Received: 7 February 2018; Accepted: 2 November 2018;
Published online 2 January 2019.

1. Camacho, A. et al. Cholera epidemic in Yemen, 2016–18: an analysis of surveillance data. *Lancet Glob. Health* **6**, e680–e690 (2018).
2. Weill, F. X. et al. Genomic history of the seventh pandemic of cholera in Africa. *Science* **358**, 785–789 (2017).
3. 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).
4. 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).
5. Mutreja, A. et al. Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477**, 462–465 (2011).
6. Naha, A. et al. Development and evaluation of a PCR assay for tracking the emergence and dissemination of Haitian variant ctxB in *Vibrio cholerae* O1 strains isolated from Kolkata, India. *J. Clin. Microbiol.* **50**, 1733–1736 (2012).
7. Chin, C. S. et al. The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364**, 33–42 (2011).
8. Domman, D. et al. Integrated view of *Vibrio cholerae* in the Americas. *Science* **358**, 789–793 (2017).
9. Katz, L. S. et al. Evolutionary dynamics of *Vibrio cholerae* O1 following a single-source introduction to Haiti. *mBio* **4**, e00398-13 (2013).
10. Ghosh Dastidar, P., Sinha, A. M., Ghosh, S. & Chatterjee, G. C. Biochemical mechanism of nitrofurantoin resistance in *Vibrio el tor*. *Folia Microbiol. (Praha)* **24**, 487–494 (1979).
11. Sandegren, L., Lindqvist, A., Kahlmeter, G. & Andersson, D. I. Nitrofurantoin resistance mechanism and fitness cost in *Escherichia coli*. *J. Antimicrob. Chemother.* **62**, 495–503 (2008).
12. Herrera, C. M. et al. The *Vibrio cholerae* VprA–VprB two-component system controls virulence through endotoxin modification. *mBio* **5**, e02283-14 (2014).
13. Matson, J. S., Livny, J. & DiRita, V. J. A putative *Vibrio cholerae* two-component system controls a conserved periplasmic protein in response to the antimicrobial peptide polymyxin B. *PLoS ONE* **12**, e0186199 (2017).
14. 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).
15. Samanta, P., Ghosh, P., Chowdhury, G., Ramamurthy, T. & Mukhopadhyay, A. K. Sensitivity to polymyxin B in El Tor *Vibrio cholerae* O1 strain, Kolkata, India. *Emerg. Infect. Dis.* **21**, 2100–2102 (2015).
16. Hasan, N. A. et al. Genomic diversity of 2010 Haitian cholera outbreak strains. *Proc. Natl Acad. Sci. USA* **109**, E2010–E2017 (2012).
17. Zarocostas, J. Cholera outbreak in Haiti-from 2010 to today. *Lancet* **389**, 2274–2275 (2017).
18. UN Office for the Coordination of Humanitarian Affairs. Humanitarian needs overview, Yemen. http://www.unocha.org/sites/unocha/files/dms/yemen_humanitarian_needs_overview_hno_2018_20171204.pdf (2017).
19. International Organization for Migration. Irregular migration in Horn of Africa increases in 2015. https://www.iom.int/news/irregular-migration-horn-africa-increases-2015 (2016).
20. Danish Refugee Council. Mixed migration in the Horn of Africa & Yemen region. *RMMS* https://reliefweb.int/sites/reliefweb.int/files/resources/RMMS%20Mixed%20Migration%20Monthly%20Summary%20September%202016_0.pdf (2016).

Acknowledgements This study was supported by the Institut Pasteur, Santé publique France, the French government's Investissement d'Avenir programme, Laboratoire d'Excellence 'Integrative Biology of Emerging Infectious Diseases' (grant number ANR-10-LABX-62-IBEID), the Wellcome Trust through grant 098051 to the Sanger Institute and the Department of Biotechnology of India. The Institut Pasteur Genomics Platform is a member of the France Génomique consortium (ANR10-INBS-09-08). We thank D. Legros, A. Fadaq, A. Alsomine, F. Bazel and H. A. Jokhdar for their support; M. Musoke and S. Vernadat for technical assistance; Z. M. Eisa for providing isolates; L. Ma, C. Fund, S. Sjunnebo and the sequencing teams at the Institut Pasteur and Wellcome Sanger Institute for sequencing the samples. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Reviewer information *Nature* thanks J. Mekalanos, C. Stine, M. Waldor and the other anonymous reviewer(s) for their contribution to the peer review of this work.

Author contributions F.-X.W. and M.-L.Q. designed the study. A.A.A., M.N., S.S.N., A.R., A.M.A., N.C.S., S.K., M.R.P., A.A., J.Y.C., J.F.W., C.S., B.B., H.H.N.A., D.K., S.M.N., M.R.M., J.K., F.J.L., A.S.A., T.R. and M.-L.Q. collected, selected and provided characterized isolates and their corresponding epidemiological information. J.R. performed the DNA extractions and phenotypic and molecular typing experiments. T.M. analysed protein data. C.B. performed the whole-genome sequencing. F.-X.W., D.D. and E.N. analysed the genomic sequencing data. F.-X.W. and D.D. wrote the manuscript, with major contributions from N.R.T. All authors contributed to the editing of the manuscript.

Competing interests The authors declare no competing interests.

Additional information
Extended data is available for this paper at https://doi.org/10.1038/s41586-018-0818-3.
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-018-0818-3.
Reprints and permissions information is available at http://www.nature.com/reprints.
Correspondence and requests for materials should be addressed to F.-X.W.
Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

METHODS

Data reporting. No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Bacterial isolates. The 116 7PET isolates sequenced in this study are listed in Supplementary Table 1 and originated from the collections of the French National Reference Centre for Vibrios and Cholera, Institut Pasteur, Paris, France ($n = 6$); the Central Public Health Laboratory of Baghdad, Iraq ($n = 11$); the Ministry of Health of South Sudan ($n = 14$); the Pasteur Institute of Iran ($n = 4$); the Maharishi Valmiki Infectious Diseases Hospital, Delhi, India ($n = 29$); the Central Public Health Laboratory of Sana'a, Yemen ($n = 39$), Amref Health Africa, Kenya ($n = 1$), the Kenya Medical Research Institute ($n = 9$) and the Ministry of Health of Saudi Arabia ($n = 3$). The isolates were characterized by standard biochemical, culture and serotyping methods[21].

Antibiotic susceptibility testing. Antibiotic susceptibility was determined by disc diffusion on Mueller–Hinton agar, in accordance with the guidelines of the Antibiogram Committee of the French Society for Microbiology[22]. The following antimicrobial drugs (Bio-Rad) were tested: ampicillin, cefalotin, cefotaxime, streptomycin, chloramphenicol, erythromycin, azithromycin, sulfonamides, trimethoprim-sulfamethoxazole, vibriostatic agent O/129, tetracycline, doxycycline, minocycline, nalidixic acid, norfloxacin, ofloxacin, pefloxacin, ciprofloxacin, nitrofurantoin, polymyxin B and colistin (polymyxin E). *E. coli* CIP 76.24 (ATCC 25922) was used as a control. The minimum inhibitory concentrations (MICs) of nalidixic acid and ciprofloxacin were determined by ETESTS (bioMérieux). The MICs of colistin and polymyxin B were determined with custom-produced Sensititre microtitre plates (ThermoFisher Scientific) and MIC test strips (Liofilchem), respectively, on 34 isolates chosen on the basis of resistance phenotype, year and country of isolation.

Total DNA extraction. Total DNA was extracted with the Wizard Genomic DNA Kit (Promega), the Maxwell 16-cell DNA purification kit (Promega) or the DNeasy Blood & Tissue Kit (Qiagen) in accordance with the manufacturer's recommendations.

Whole-genome sequencing. High-throughput genome sequencing was carried out at the genomics platform of Institut Pasteur ($n = 107$) or at the Wellcome Sanger Institute ($n = 9$) on Illumina platforms generating 92–295-bp paired-end reads, yielding a mean of 117-fold coverage (minimum 13.5-fold, maximum 639-fold). Short-read sequence data were submitted to the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena), under study accession numbers PRJEB24611 and ERP021285 and the genome accession numbers are provided in Supplementary Table 1.

Genomic sequence analyses. The genomic sequences were processed and analysed as previously described[2]. In brief, for each sample, sequence reads were mapped against reference genome *V. cholerae* O1 El Tor N16961 (GenBank accession numbers LT907989 and LT907990) using SMALT version 0.7.4 (http://www.sanger.ac.uk/science/tools/smalt-0) to produce a BAM file. Variants were detected with samtools mpileup[23] version 0.1.19 with parameters '-d 1000 –DsugBf' and bcftools[23] version 0.1.19 to produce a BCF file of all variant sites. The bcftools variant quality score had to be greater than 50 (quality > 50) and mapping quality greater than 30 (map quality > 30). The majority base call was required to be present in at least 75% of reads mapping to the base (ratio ≥ 0.75) and the minimum mapping depth required was four reads, at least two of which had to map to each strand (depth ≥ 4, depth strand ≥ 2). A pseudogenome for each sample was constructed by substituting the base call at each site (variant and non-variant) in the BCF file in the reference genome. While this paper was under review, another paper[4] was published that included three genome sequences from Ugandan isolates that belonged to the T13 sublineage. These three genome sequences were available as contig files and were added to the alignment with Snippy version 4.1.0 (https://github.com/tseemann/snippy), using the '–ctgs' flag to call SNVs between the contigs and the reference genome. Short reads were assembled with SPAdes[24] version 3.8.2 and annotated with Prokka[25] version 1.5.

The code for the pipelines from the Sanger Institute used can be found here: https://github.com/sanger-pathogens/vr-codebase.

Phylogenetic analysis. Repetitive (insertion sequences and the TLC-RS1-CTX region) and recombinogenic (VSP-II) regions were masked from the alignment[2]. Putative recombinogenic regions were detected and masked with Gubbins[26] version 1.4.10. A maximum-likelihood phylogenetic tree was built from an alignment of 9,986 chromosomal SNVs, with RAxML[27] version 8.2.8 under the GTR model with 100 bootstraps.

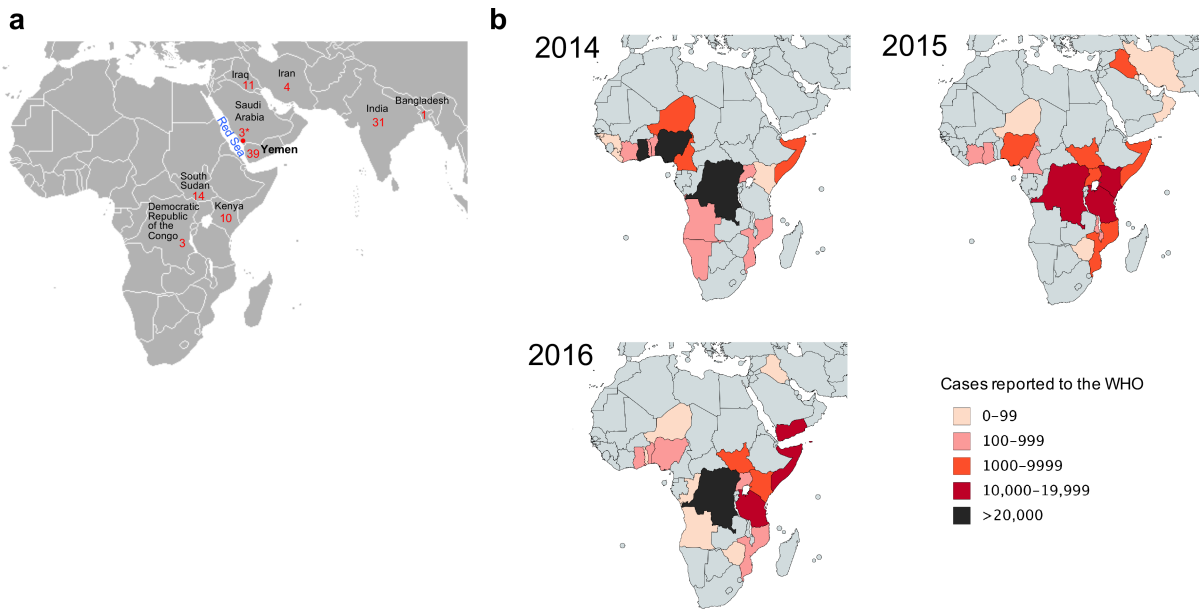
BEAST[28] version 1.10.1 was used to estimate time-resolved phylogenies for a spatially and temporally representative subset of 81 7PET isolates under the GTR nucleotide substitution model. We tested a combination of molecular clock and tree prior models to identify the best fit (Extended Data Table 1). Both path and stepping-stone sampling showed the best fit to be an uncorrelated relaxed clock (lognormal distribution of rates) model with a Bayesian skyline coalescent tree prior. Priors were kept at default values, with the exception of the 'constant.popSize' value, which was set to a lognormal distribution (initial value = 1, mu = 1, sigma = 10) under the constant population coalescence tree prior. The choice of model had little influence on the dating of key nodes in this analysis (Extended Data Table 1). For each model, we ran three independent Markov chain Monte Carlo chains over 50 million steps, sampling every 2,000 steps. We used a burn-in of 5 million steps for each chain and then combined chains, resampling every 10,000 steps. The effective sample size for all estimated parameters was greater than 200. We tested for an adequate temporal signal, using TempEst[29] version 1.5, by calculating the linear regression between the root-to-tip distance and isolation date for each sample. We also performed 20 date-randomization tests with the R package TipDatingBeast[30] to assess the mean rate under the uncorrelated lognormal relaxed molecular clock (ucld.mean parameter).

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

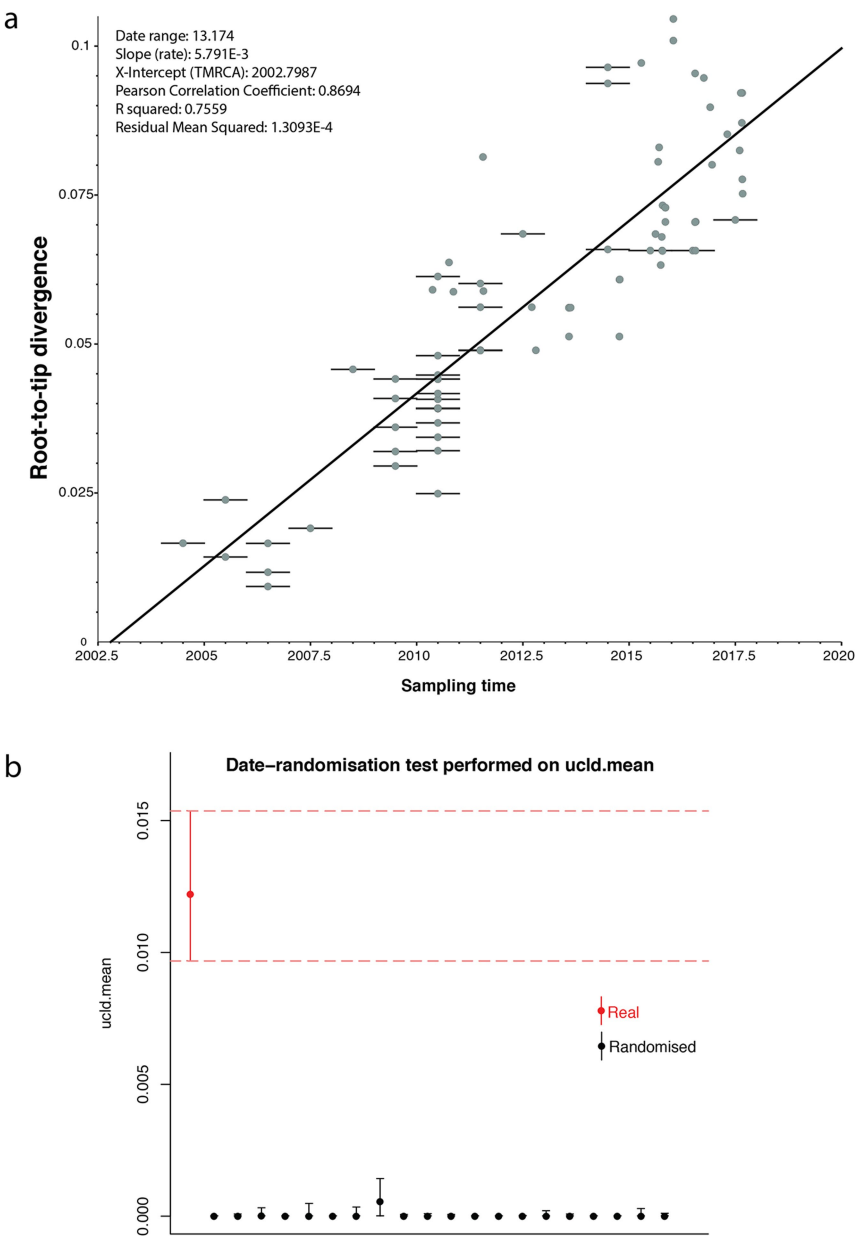
Data availability

The whole-genome alignment for the 1,203 genomes and other files that support the findings of this study have been deposited in FigShare: https://figshare.com/s/b70a9efac9cf2625480e. Short-read sequence data were submitted to the ENA, under study accession numbers PRJEB24611 and ERP021285 and the genome accession numbers are provided in Supplementary Table 1. Phylogeny and metadata can be viewed interactively at https://microreact.org/project/globalcholera.

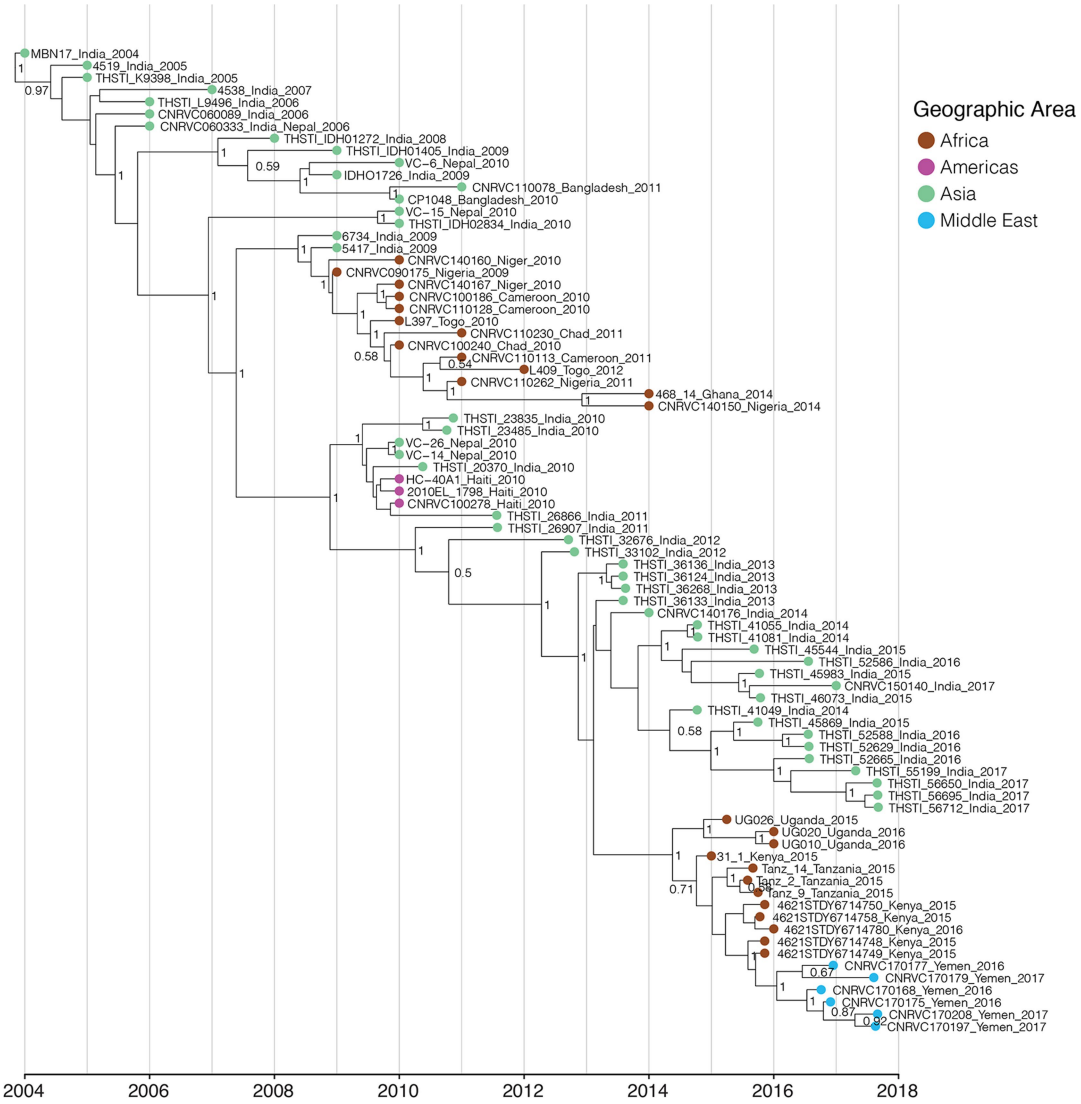
21. Dodin, A. & Fournier, J. M. *Laboratory Methods for the Diagnosis of Cholera Vibrio and Other Vibrios* 59–82 (Institut Pasteur Paris, Paris 1992).
22. CA-SFM & EUCAST. *Comité de l'Antibiogramme de la Société Française de Microbiologie Recommandations 2017*. http://www.sfm-microbiologie.org/UserFiles/files/casfm/CASFMV1_0_MARS_2017.pdf (2017).
23. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
24. Bankevich, A. et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477 (2012).
25. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068–2069 (2014).
26. Croucher, N. J. et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43**, e15 (2015).
27. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
28. Drummond, A. J., Suchard, M. A., Xie, D. & Rambaut, A. Bayesian phylogenetics with BEAUti and the BEAST 1.7. *Mol. Biol. Evol.* **29**, 1969–1973 (2012).
29. Rambaut, A., Lam, T. T., Max Carvalho, L. & Pybus, O. G. Exploring the temporal structure of heterochronous sequences using TempEst (formerly Path-O-Gen). *Virus Evol.* **2**, vew007 (2016).
30. Rieux, A. & Khatchikian, C. E. tipdatingbeast: an R package to assist the implementation of phylogenetic tip-dating tests using beast. *Mol. Ecol. Resour.* **17**, 608–613 (2017).



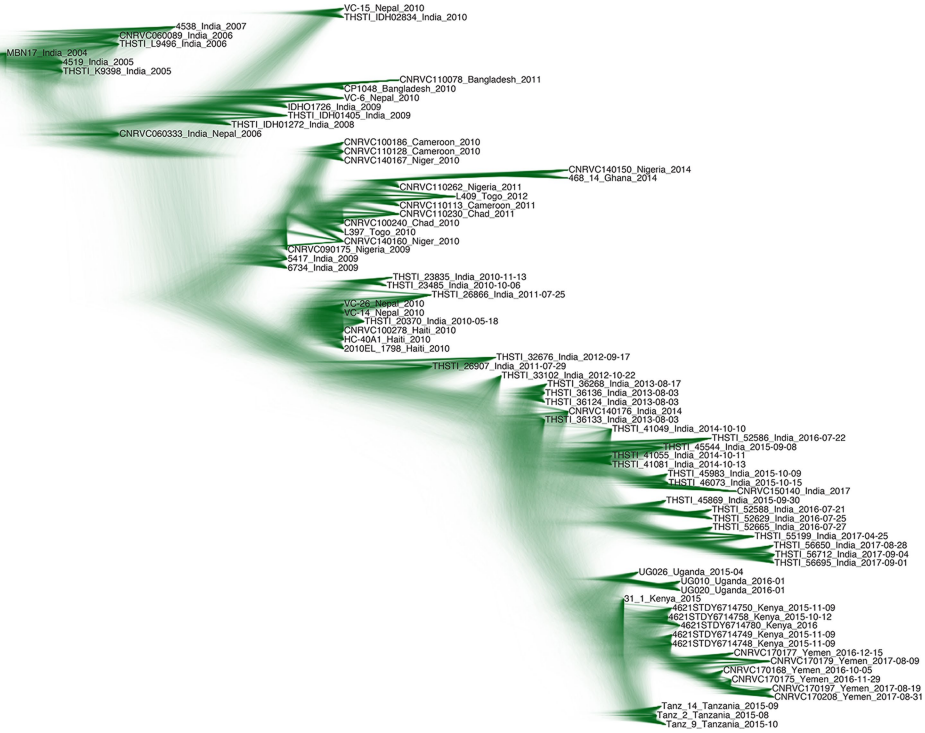
Extended Data Fig. 1 | Geographic location of the sequenced *V. cholerae* O1 El Tor isolates and number of reported cholera cases. a, Geographic location of the 116 *V. cholerae* O1 El Tor isolates sequenced. The number of isolates collected per country is indicated. The three isolates collected in Jizan, Saudi Arabia (denoted by an asterisk) were from Yemeni refugees originating from Hajjah District. The map is a cropped version of the one available at https://commons.wikimedia.org/wiki/File:BlankMap-World.png. **b**, Number of cholera cases per country reported to the WHO (World Health Organisation) between 2014 and 2016. The total number of cholera cases reported to the WHO by the countries was 268,337. The maps were created using Paintmaps, a free online map generating tool (http://www.paintmaps.com/).



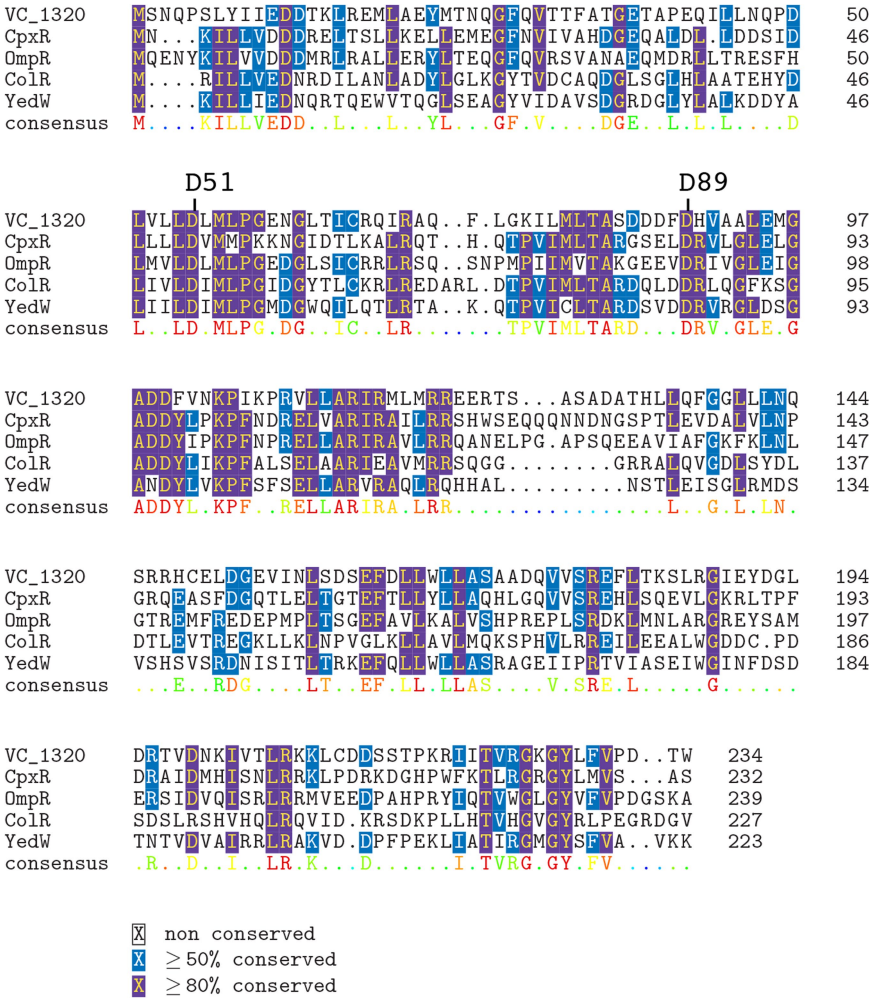
Extended Data Fig. 2 | Assessment of the temporal signal within the dataset. a, Linear regression of the root-to-tip distance against sampling time obtained with TempEst[29] using a maximum-likelihood phylogeny of 81 representative seventh pandemic *V. cholerae* O1 isolates (that is, those used for the BEAST analysis). Bars on nodes indicate the precision of the isolation date (for example, if only the year of isolation is known, the bar spans the entire year). **b**, Comparison of the ucld.mean parameter estimated from 20 date-randomization BEAST experiments and the original dataset. The rate for the correctly dated tree is shown in red. The median and 95% Bayesian credible interval for the ucld.mean parameter are provided.



Extended Data Fig. 3 | Timed phylogeny of the *ctxB7* clade. Maximum clade credibility tree produced with BEAST[28] for a subset of 81 representative isolates of the distal part of the genomic wave 3 (that is, those with the *ctxB7* allele). The nodes supported by posterior probability values ≥0.5 are indicated.



Extended Data Fig. 4 | Visualization of the posterior distribution of trees from the BEAST Markov chain Monte Carlo analysis. The opacity of the branches is scaled according to the number of times a clade is seen in the distribution. There is high support for the East Africa/Yemen clade. The uncertainty in the placement of the node for the Indian/East African isolates is the reason for the low posterior support value for this node in Extended Data Fig. 3.



Extended Data Fig. 5 | Multiple sequence alignment of VprA (VC1320) with two-component response regulators. A non-synonymous mutation at position 89 of VC1320 that resulted in a D-to-N amino acid change was associated with a phenotype of polymyxin B susceptibility.

Extended Data Table 1 | Summary of the Bayesian models used for BEAST[28] analyses

	Yemeni isolates tMRCA			African/Yemeni isolates tMRCA		
Model	**Median**	**Upper 95% HPD**	**Lower 95% HPD**	**Median**	**Upper 95% HPD**	**Lower 95% HPD**
Strict, Constant	2016.04	2016.34	2015.72	2014.14	2014.62	2013.59
Strict, Bayesian skyline	2016.03	2016.35	2015.71	2014.16	2014.63	2013.61
UCLN, Bayesian skyline	2016.05	2016.44	2015.67	2014.38	2014.86	2013.74

Model	log MLE Path Sampling	Rank	log MLE Stepping Stone Sampling	Rank
Strict, Constant	-5481314.81	3	-5481312.99	2
Strict, Bayesian skyline	-5481314.47	2	-5481313.00	3
UCLN, Bayesian skyline	-5481313.28	1	-5481311.50	1

Analyses were carried out on a subset of 81 representative *V. cholerae* O1 isolates of the distal part of the genomic wave 3. HPD, highest posterior density region; MLE, marginal likelihood estimate; tMRCA, time to the most recent common ancestor; UCLN, uncorrelated lognormal relaxed clock.

Extended Data Table 2 | Gene alteration frequencies in isolates susceptible or resistant to certain antibiotics

	Number (%) of isolates:		
VCA_0637	FT^S	FT^R	
Wild-type	258 (94.2)	0 (0)	
Genetic alteration	16 (5.8)	446 (100)	
Total	**274**	**446**	**720***

	Number (%) of isolates:		
VC_0715	FT^S	FT^R	
Wild-type	264 (96.4)	5 (1.1)	
Genetic alteration	10 (3.6)	441 (98.9)	
Total	**274**	**446**	**720***

	Number (%) of isolates:		
VCA_0637 and VC_0715	FT^S	FT^R	
Wild-type	258 (94.2)	0 (0)	
Genetic alteration	16 (5.8)	446 (100)	
Total	**274**	**446**	**720***

	Number (%) of isolates:		
VC_1320	POL^S	POL^R	
Wild-type	2 (3.1)	40 (97.6)	
D89N	63 (96.9)	1 (2.4)	
Total	**65**	**41**	**106****

FT, nitrofurantoin; POL, polymyxin B.
*720 genomes with antimicrobial susceptibility testing data (this study and published previously[2]).
**106 genomes with antimicrobial susceptibility testing data (this study).

nature research

Corresponding author(s): WEILL, Nature 2017-12-16825

Reporting Summary

Nature Research 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 Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. 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
☒	☐	An indication of 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 statistics 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
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Excel version 15.41
Data analysis	SMALT version 0.7.4, samtools mpileup version 0.1.19, bcftools version 0.1.19, Gubbins version 1.4.10, RAxML version 7.8.6, BEAST version 1.10.1, FigTree version 1.4.2, LogCombiner version 1.7.5, TreeAnnotator version 1.7.5, SPAdes version 3.8.2, TempEst version 1.5, R package TipDatingBeast, Snippy version 4.1.0

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/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

Short-read sequence data were submitted to the European Nucleotide Archive (ENA) (http://www.ebi.ac.uk/ena), under study accession numbers PRJEB24611 and

ERP021285, and the genome accession numbers are provided in Supplementary Table 1. The whole-genome alignment for the 1203 genomes and other files that support the findings of this study have been deposited in FigShare: https://figshare.com/s/b70a9efac9cf2625480e

Field-specific reporting

Please select 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/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

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

Sample size	We have sequenced all the 45 Yemeni V. cholerae isolates sent by two different laboratories in Yemen and Saudi Arabia.
Data exclusions	Three genomes out of 45 were excluded due to low coverage or mixup determined by the quality control step. The threshold for an appropriate coverage was predefined (15X or more). The mixup was discovered by analysing a discrepancy between phenotypic testing and genome content.
Replication	The Yemeni genomes have been analysed by two different phylogenetic approaches (Maximum Likelihood and Bayesian). This was replicated two times during the revision process by including 14 new genomic sequences then by including three genomes. All attempts at replication were successful as we have obtained a same tree topology.
Randomization	It is not applicable as for the phylogenetic tree we have analysed all the genomes that have passed the sequencing quality control (42/45).
Blinding	The phylogenetic trees were initially drawn without any geographic information associated with the genomes.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

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