Supplementary Materials for

# Defining endemic cholera at three levels of resolution within Bangladesh

Daryl Domman[1‡], Fahima Chowdhury[2], Ashraf I. Khan[2], Matthew J. Dorman[1], Ankur Mutreja[1,3], Muhammad Ikhtear Uddin[2], Anik Paul[2], Yasmin A. Begum[2], Richelle C. Charles[4,5], Stephen B. Calderwood[4,5], Taufiqur R. Bhuiyan[2], Jason B. Harris[4,5,6], Regina C. LaRocque[4,5], Edward T. Ryan[4,5,7*], Firdausi Qadri[2*], Nicholas R. Thomson[1,8*‡]

**This PDF file includes:**

## Methods

### Bacterial isolates

*Vibrio cholerae* O1 and O139 isolates, as well as household contact data in this analysis, build upon a study conducted in Dhaka, Bangladesh between March 2002 and June 2005[4], in which the strains in this study were analysed using variable number tandem repeat analysis[4]. Briefly, patients with acute watery diarrhea and culture-confirmed *V. cholerae* O1 or O139 presenting to the International Centre for Diarrhoeal Disease Research in Dhaka, Bangladesh (icddr,b) were enrolled as cholera index patients following an informed consent process. Willing household contacts of these cholera index patients were also enrolled. Household contacts were defined as individuals sharing a cooking pot with the patient for at least three or more days prior to enrolment of the index patient. Household contacts were monitored for three weeks from the time of enrolment, and daily rectal swabs were collected for culture on days 1-6, 13, and 20 following the index patient's presentation at icddr,b. In total, this study involved members of 103 households and index patients, and included 224 individuals (103 index patients and 121 household contacts). In total, 303 *V. cholerae* isolates were acquired from these individuals. More than one individual sample was obtained from 79 households, and *V. cholerae* was isolated from 45 individuals on more than one day while they were enrolled in the study. Five technical replicates were sequenced from four patients in order to characterize SNP artifacts from culturing procedures and sequencing. Microbiological analyses, serogrouping and serotyping, were performed as previously described[4]. Vibriocidal titers were assessed as previously described[8]. Participants provided written informed consent, and the study was approved by the Institutional Review Board (IRB) in both Dhaka and at Massachusetts General Hospital.

### Whole genome sequencing

Whole genome sequencing was performed at the Wellcome Trust Sanger Institute using the Illumina HiSeq and MiSeq platforms to generate 100-350 bp paired-end reads. Short read data have been deposited in the European Nucleotide Archive (ENA) database under the accession codes provided in Supplementary Table 2. Short reads were assembled using Spades v3.8.2[19] and annotated using Prokka v1.5[20] as part of a high-throughput improvement pipeline[21]. Genome completeness estimates and checks for contamination were performed using CheckM [22].

### Additional genomes

Previously published genomes used in these analyses are listed in Supplementary Table 2.

**Read alignment and detection of SNPs**

A reference based alignment for the 7th pandemic *V. cholerae* (7PET) isolates was obtained by mapping paired-end Illumina reads for each of the 813 isolates to the *Vibrio cholerae* O1 El Tor reference N16961 (NCBI accession numbers LT907989/LT907990) using SMALT v.0.7.4 (http://www.sanger.ac.uk/science/tools/smalt-0). Variant detection was performed using samtools mpileup v0.1.19[23] with parameters "-d 1000 -DSugBf" and bcftools v0.1.19 to produce a BCF file of all variant sites. High quality SNPs were determined as previously described[24]. Putative recombinant regions (see below) were detected and filtered from the alignment using Gubbins[25], resulting in a final alignment of 6,241 variable (SNP) sites. The alignment of the 300 7PET isolates from this study was performed as above, but yielded a final alignment of 533 variable (SNP) sites after putative recombination sites were removed.

**Detection of recombinant sites**

To determine the potential impact of recombination within this dataset, we also ran ClonalFrameML[26]. The ClonalFrameML analysis revealed an overall r/m = 0.47, with the ratio of recombination to mutation rates (R/Θ) = 0.0439438, the mean length of recombination events ($\delta$) = 261.217 bp, and the average distance between events ($\nu$) = 0.041323. Gubbins reports r/m values per branch. Of the 600 nodes within the phylogenetic tree (tips plus internal nodes), only six nodes have an r/m > 0, varying from 0.186 to 4.76. The largest value is reported for the basal node for all O139 isolates. The region identified as recombinogenic is that which encodes the O-antigen biosynthesis genes, which is to be expected.

**Phylogenetic analysis**

A Bayesian phylogenetic reconstruction of 7PET isolate genomes circulating in Dhaka (n = 300) was performed using PhyloBayes v3[27] under the GTR model. Two independent chains were run in parallel, and the *bpcomp* and *tracecomp* programs were used to assess convergence. We judged that analyses had converged when the maximum discrepancies in bipartition frequencies (*bpcomp*) and summary statistics (*tracecomp*) between the two chains had dropped below 0.1, and when the effective sample size of each parameter was at least 100. Maximum likelihood (ML) phylogenetic trees were constructed using RAxML v8.2.8[28] and IQ-Tree[29] under the GTR model with the gamma distribution to model site heterogeneity (GTRGAMMA), with 500 bootstrap replicates. The ML analysis for the 7PET dataset (813 isolates) was performed using RAxML on an alignment of 6,241 variable sites with putative recombinant sites removed described above. Phylogenetic trees and networks were visualized with ggtree[30] and iTOL[31]. Networks were created by drawing an arc between

each isolate within a household, anchored to the index case. For households with no index case reported, the first case reported in the household was used. Roary[32] was used to generate a core gene alignment of 1,933 genes for the non-7PET dataset. FastTree[33] was used to generate a maximum-likelihood phylogeny using the variable site alignment (165,443 sites) generated by SNP-sites v.2.3.2[34].

**Temporal analysis**

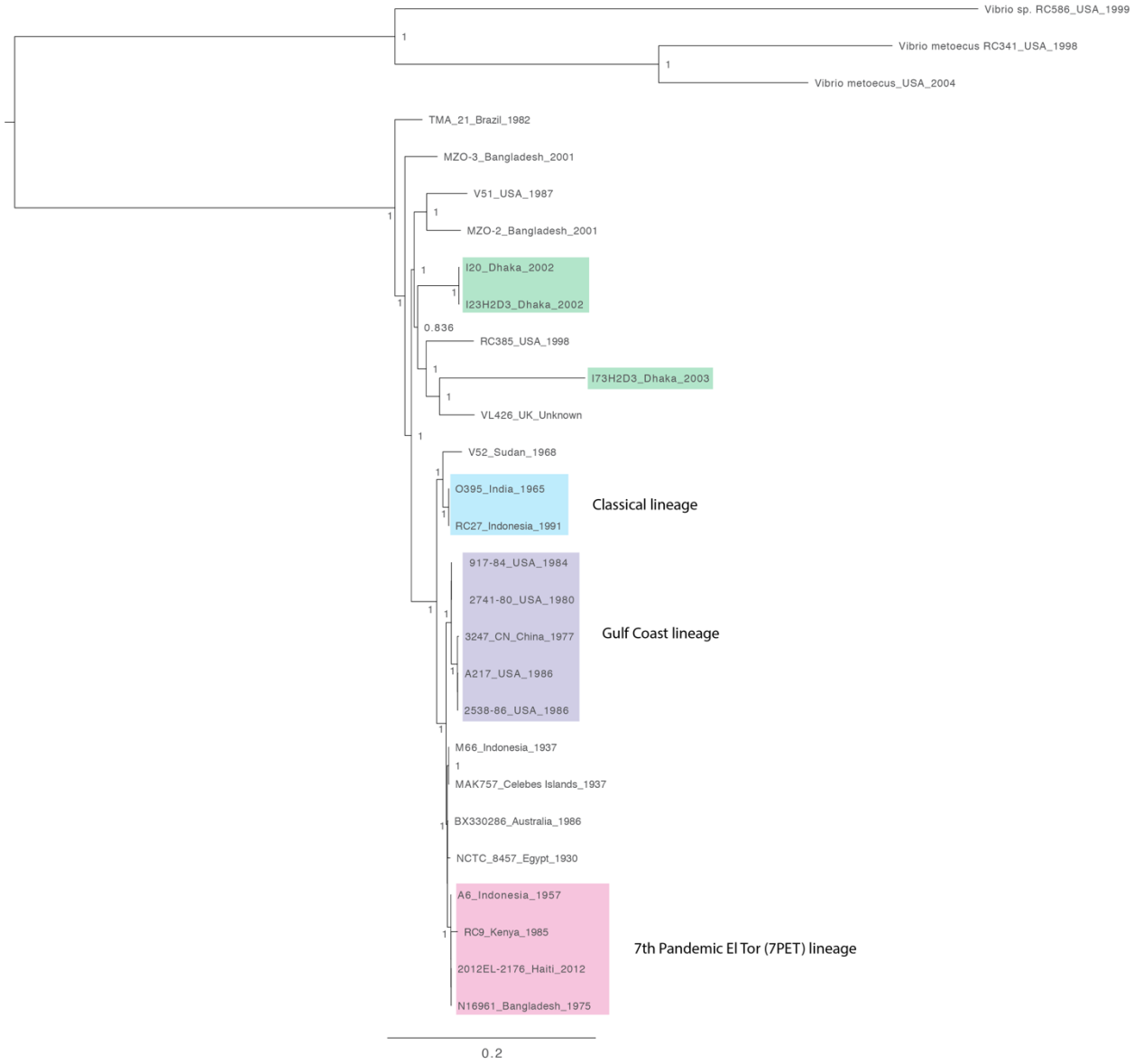
The temporal signal within the dataset was investigated within the ML tree using TempEst v.1.5[35] by calculating a linear regression between the root-to-tip distance and the year of isolation for each sample. Randomization of the dates of isolation removed any correlation. A time-scaled phylogeny was produced by applying the LSD algorithm[36], which can scale to a large number of taxa. Here, we applied the genome-wide mutation rate obtained by Mutreja *et al*.[37] ($8.3 \times 10^{-7}$ substitutions $site^{-1}$ $year^{-1}$). LSD version 0.3beta was run with constrained mode ("-c"), using variances from the estimated branch lengths ("-v 2"), and confidence intervals computed from 1,000 simulated trees ("-f 1000"). The time to most recent common ancestor (tMRCA) was determined to be 1909. (1909.737, 95% CI: 1906.081, 1913.596) for all 7PET.

**Comparative genomic analyses**

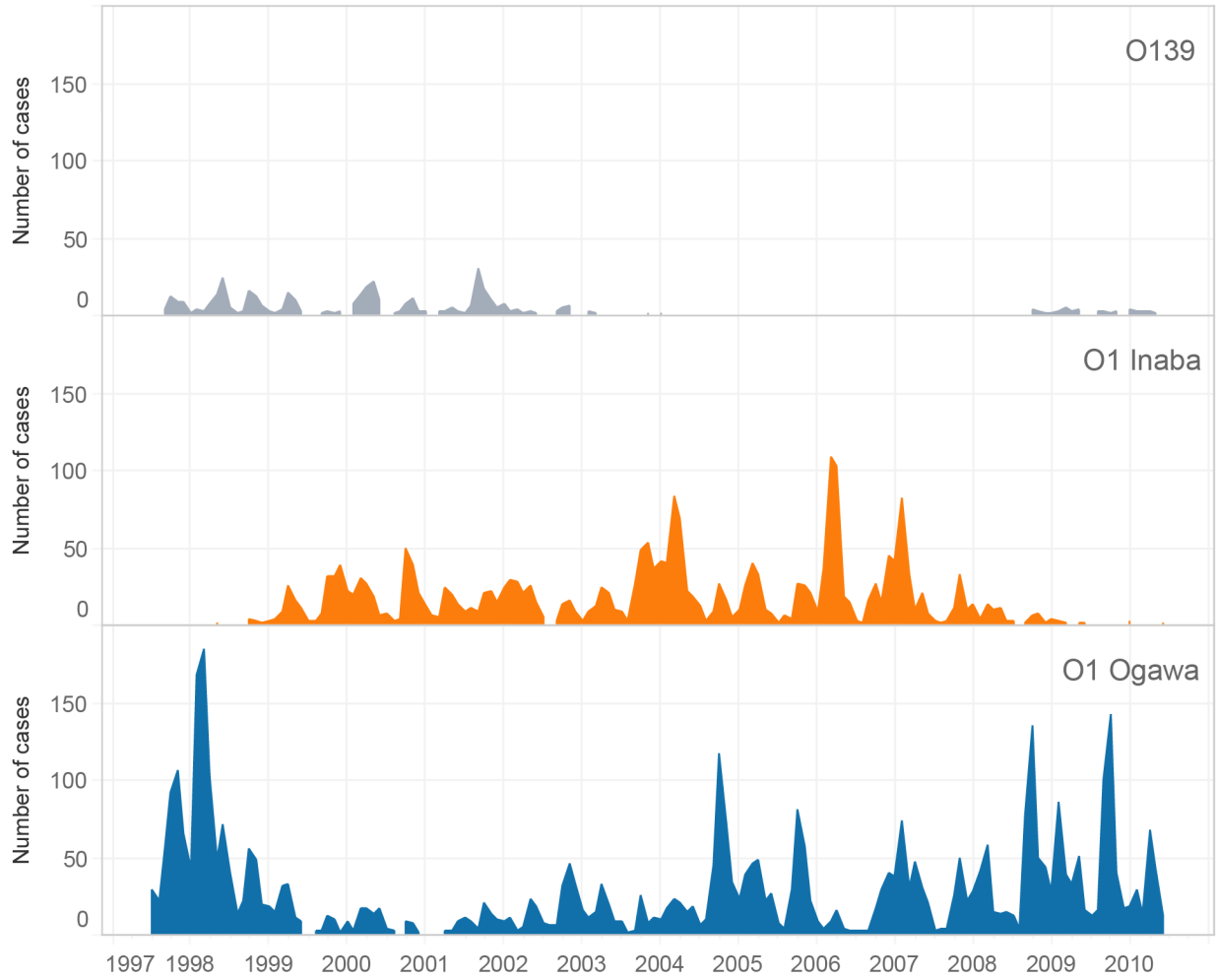
For *in silico* typing of the 7PET Ogawa/Inaba serotypes, we used ARIBA[38] to call variants against the wild type Ogawa *V. cholerae* V06-92 genome sequence (ENA accession AEN80191). Similarly, ARIBA was used to call variants of the *ctxB* allele against the *V. cholerae* N16961 El Tor sequence (NCBI accession numbers LT907989/LT907990).

**Geospatial and statistical analyses**

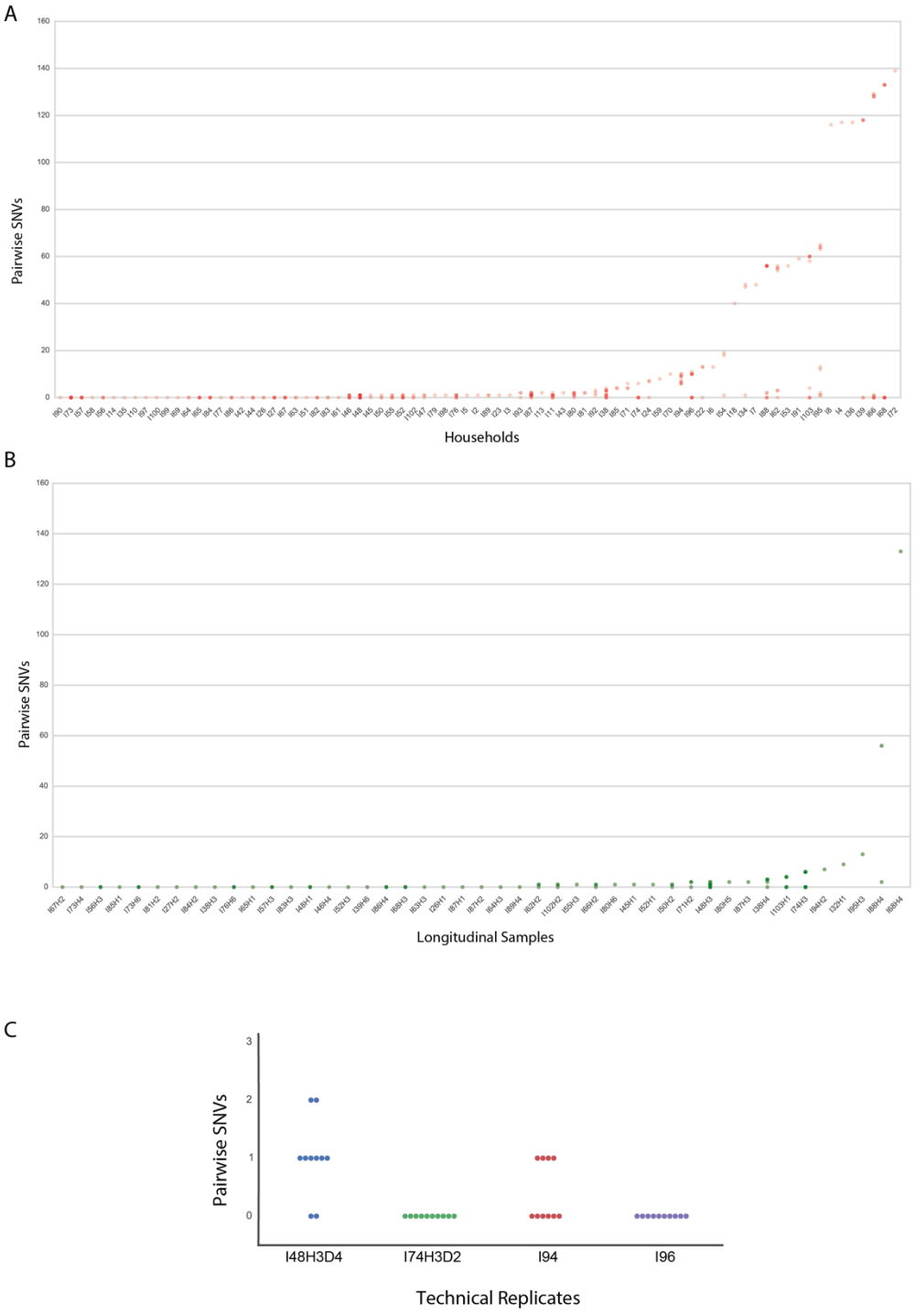
Pairwise distances between isolates were computed using the Python package *GeoPy* based on geospatial data (Supplementary Table 1). The map used for Figure 2 was created using MapBox (http://mapbox.com/), and data were visualized using QGIS v2.16 (http://www.qgis.org/). All statistical tests were performed in Python using the *pandas, numpy,* and *scipy* packages. All source code is available upon request.



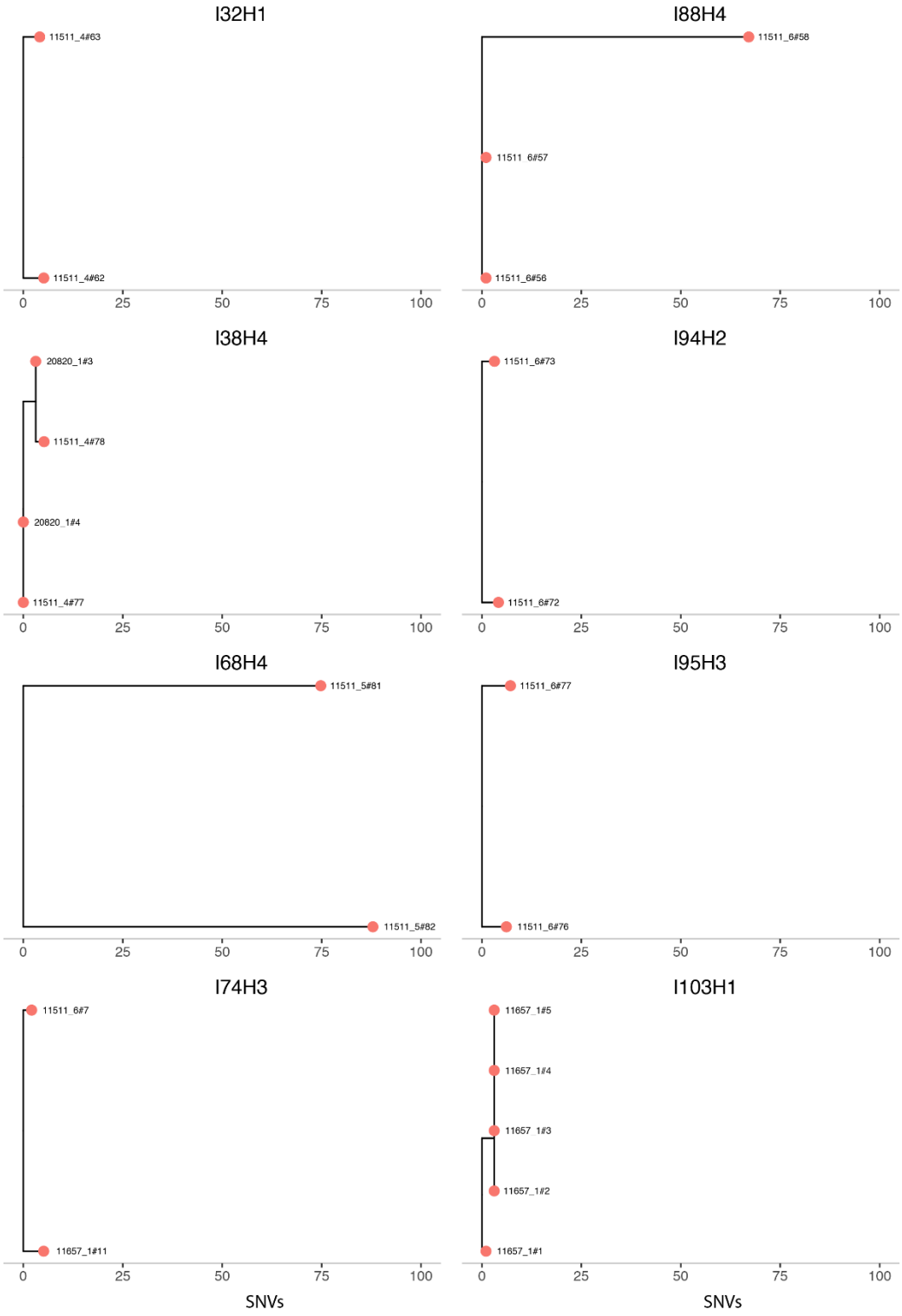

**Supplementary Figure 1**. **Maximum likelihood phylogeny showing the relationships between *Vibrio cholerae*.** Three isolates sampled in Dhaka, shown in green, do not belong to the 7$^{th}$ Pandemic El Tor (7PET) lineage. The location and date of isolation for each isolate is listed. *V. metoecus* and *Vibrio sp.* RC586 were used as outgroups for the phylogeny.



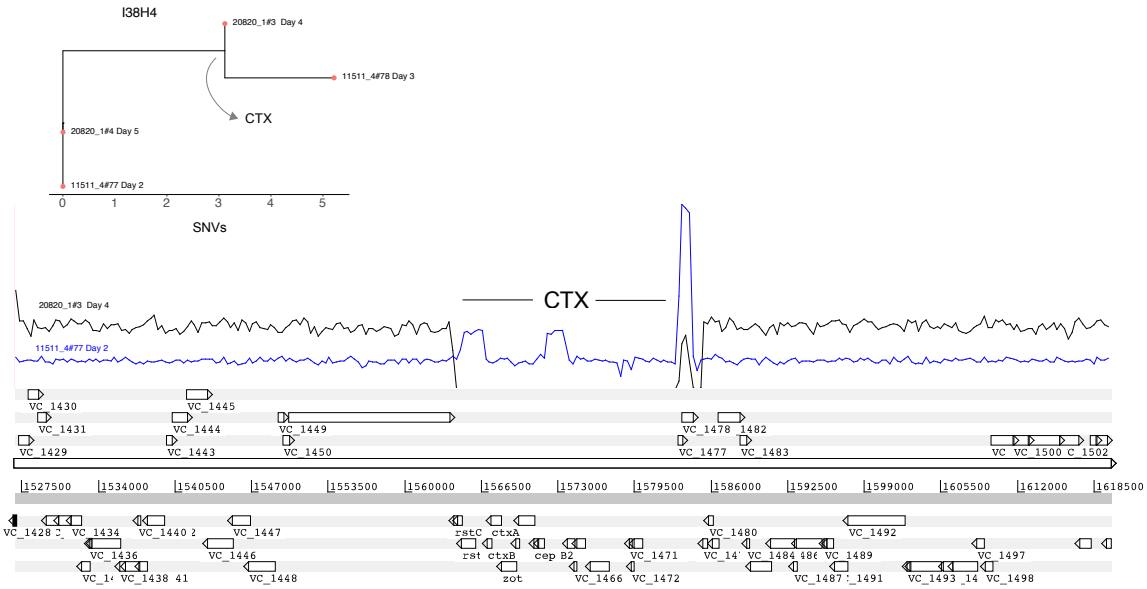

**Supplementary Figure 2**. **Cholera incidence from icddr,b hospital in Dhaka, Bangladesh.** The diarrheal disease surveillance system at icddr,b enrolls every 50$^{th}$ individual for full analysis. The different panels discriminate between O1 serotypes and the O139 serogroup.



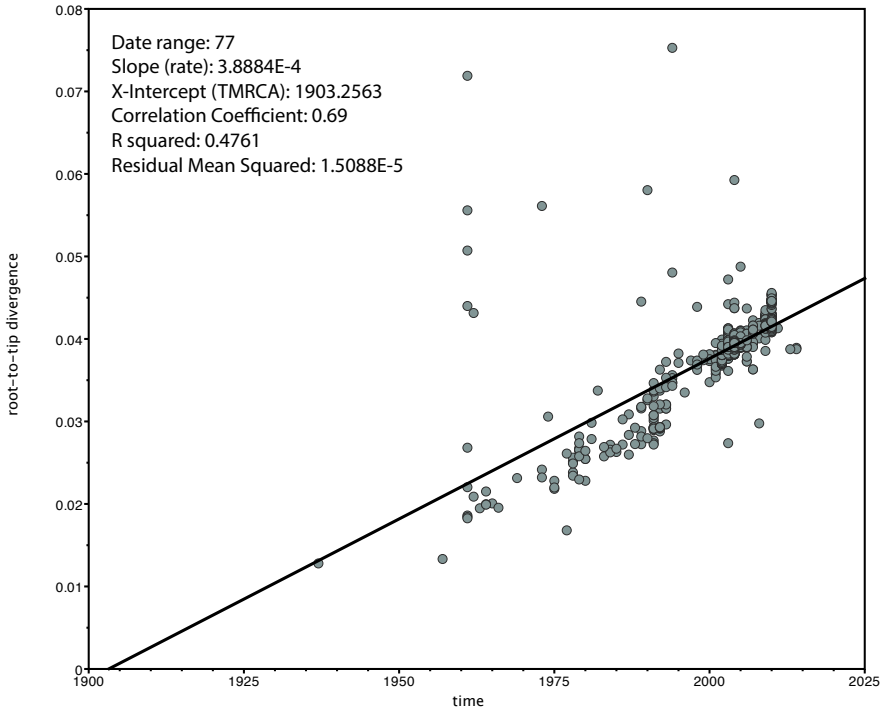

**Supplementary Figure 3**. **Distribution of SNVs across households and individuals**. **(A)** The pairwise comparison of SNVs shared across households is ordered from least to greatest variation within a single household. **(B)** Pairwise variation across individuals sampled more than once. **(C)** Pairwise variability within technical replicates.



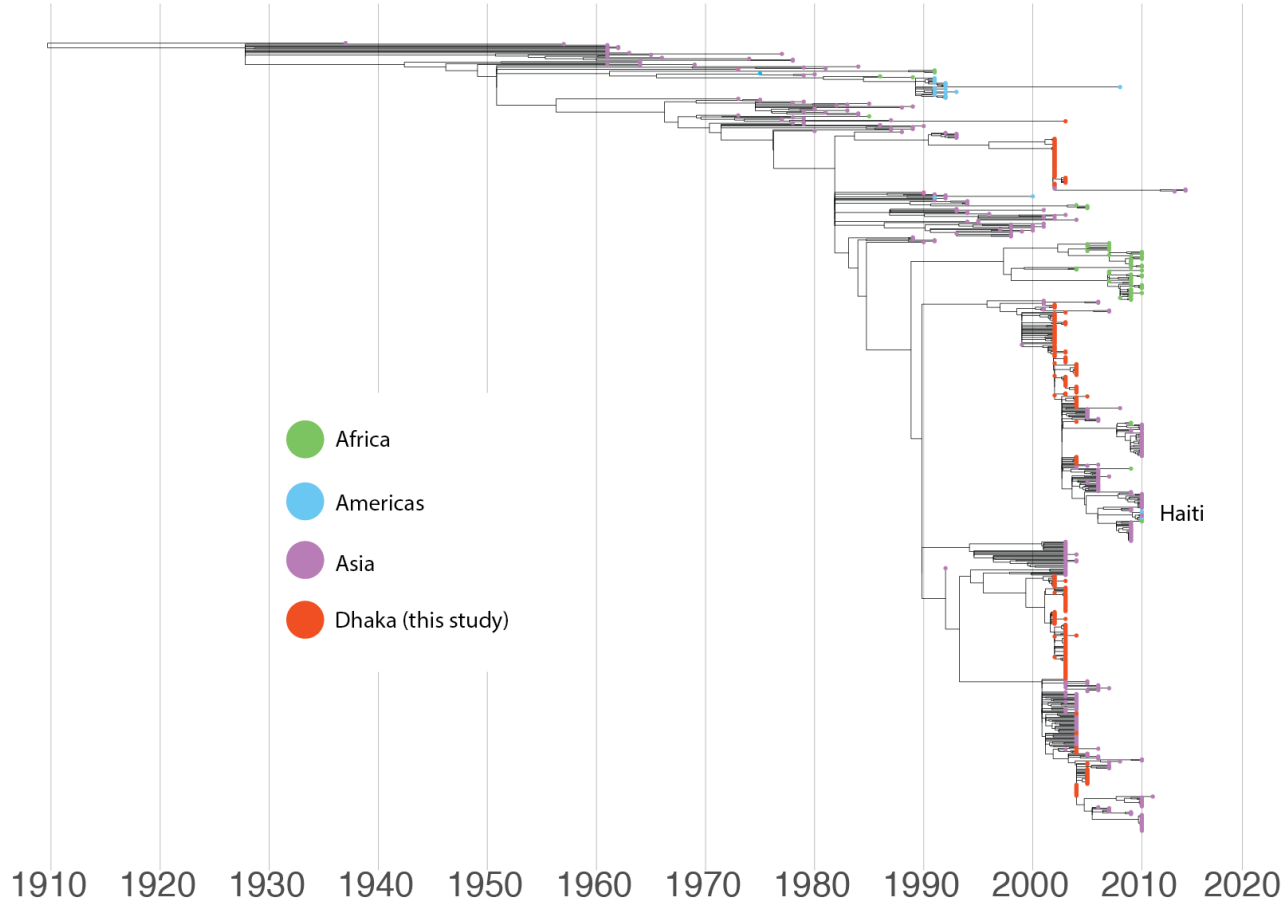

**Supplementary Figure 4. Phylogenies of isolates sampled from individuals over the course of an infection.** Each panel depicts the relatedness of samples from the same individual. The scale is number of SNVs per site.




**Supplementary Figure 5. Loss of CTX bacteriophage within an individual.** The phylogenetic relatedness of the isolates from this individual is shown in the top diagram. The bottom panel shows the coverage of reads mapped to the CTXφ region of the reference genome N16961 for samples from Day 2 and Day 4.




**Supplementary Figure 6**. **Temporal signal within the 813 7PET genomes**. Regression of the year of isolation versus root-to-tip divergence derived from the maximum likelihood tree of the 7PET lineage in Fig 3. The hypermutator strains (n = 11) described by Didelot *et al*.[39] contribute the majority (11 of 13) of the outliers seen in the root-to-tip regression.




**Supplementary Figure 7. Time-scaled phylogeny for the 7PET *V. cholerae* lineage.** The tips are colored according to the geographic origin of the isolates. The nodes are in the same order as in Figure 5.

References:


19. Bankevich, A. *et al.* SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *J. Comput. Biol.* **19,** 455–477 (2012).
20. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinforma. Oxf. Engl.* **30,** 2068–2069 (2014).
21. Page, A. J. *et al.* Robust high-throughput prokaryote de novo assembly and improvement pipeline for Illumina data. *Microb. Genomics* **2,** (2016).
22. Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* **25,** 1043–1055 (2015).
23. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).
24. Wong, V. K. *et al.* Phylogeographical analysis of the dominant multidrug-resistant H58 clade of Salmonella Typhi identifies inter- and intracontinental transmission events. *Nat. Genet.* **47,** 632–639 (2015).
25. Croucher, N. J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res.* **43,** e15–e15 (2015).
26. Didelot, X. & Wilson, D. J. ClonalFrameML: Efficient Inference of Recombination in Whole Bacterial Genomes. *PLOS Comput. Biol.* **11,** e1004041 (2015).
27. Lartillot, N., Lepage, T. & Blanquart, S. PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* **25,** 2286–2288 (2009).
28. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30,** 1312–1313 (2014).
29. Nguyen, L.-T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol. Biol. Evol.* **32,** 268–274 (2015).

30. Yu, G., Smith, D. K., Zhu, H., Guan, Y. & Lam, T. T.-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol. Evol.* **8,** 28–36 (2017).

31. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44,** W242-245 (2016).

32. Page, A. J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31,** 3691–3693 (2015).

33. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS ONE* **5,** (2010).

34. Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb. Genomics* **2,** (2016).

35. 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,** (2016).

36. To, T.-H., Jung, M., Lycett, S. & Gascuel, O. Fast Dating Using Least-Squares Criteria and Algorithms. *Syst. Biol.* **65,** 82–97 (2016).

37. Mutreja, A. *et al.* Evidence for several waves of global transmission in the seventh cholera pandemic. *Nature* **477,** 462–465 (2011).

38. Hunt, M. *et al.* ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *bioRxiv* 118000 (2017). doi:10.1101/118000

39. Didelot, X. *et al.* The Role of China in the Global Spread of the Current Cholera Pandemic. *PLoS Genet.* **11,** (2015).