BRIEF REPORT

Second-Pandemic Strain of *Vibrio cholerae* from the Philadelphia Cholera Outbreak of 1849

Alison M. Devault, M.A., G. Brian Golding, Ph.D., Nicholas Waglechner, M.Sc., Jacob M. Enk, M.Sc., Melanie Kuch, M.Sc., Joseph H. Tien, Ph.D., Mang Shi, M.Phil., David N. Fisman, M.D., M.P.H., Anna N. Dhody, M.F.S., Stephen Forrest, M.Sc., Kirsten I. Bos, Ph.D., David J.D. Earn, Ph.D., Edward C. Holmes, Ph.D., and Hendrik N. Poinar, Ph.D.

SUMMARY

In the 19th century, there were several major cholera pandemics in the Indian subcontinent, Europe, and North America. The causes of these outbreaks and the genomic strain identities remain a mystery. We used targeted high-throughput sequencing to reconstruct the *Vibrio cholerae* genome from the preserved intestine of a victim of the 1849 cholera outbreak in Philadelphia, part of the second cholera pandemic. This O1 biotype strain has 95 to 97% similarity with the classical O395 genome, differing by 203 single-nucleotide polymorphisms (SNPs), lacking three genomic islands, and probably having one or more tandem cholera toxin prophage (CTX) arrays, which potentially affected its virulence. This result highlights archived medical remains as a potential resource for investigations into the genomic origins of past pandemics.

CHOLERA IS A DIARRHEAL DISEASE CAUSED BY COLONIZATION OF THE INtestines by cholera toxin–expressing strains of the waterborne enteric bacterium *V. cholerae*. An outbreak can arise suddenly, especially in vulnerable populations with compromised sanitation infrastructure, as in the devastating 2010 outbreak in Haiti.[1] In 2012 alone, *V. cholerae* infected 3 million to 4 million people, killing nearly 100,000.[2] Although all pathogenic *V. cholerae* strains possess a similar genomic backbone that may have facilitated adaptation to human intestinal mucosa,[3,4] the predominant pathogenic strain, serogroup O1, harbors two genetically distinct biotypes: classical and El Tor (for descriptions of these and other terms, see the Glossary). In the 20th century, for unknown reasons, El Tor replaced classical as the dominant biotype. There have been seven documented pandemics since 1817,[5] but the causal *V. cholerae* strains have been genetically characterized only for the two most recent outbreaks. Therefore, although most assume that the classical biotype was responsible for the earlier pandemics,[6] the identities of the strains that caused them remain unknown.

Diverse tissue specimens archived by medical practitioners at the time of an outbreak represent an essentially untapped genetic museum for pathogen research. One extraordinary collection includes a preserved intestine from a patient who died from cholera in the 1849 Philadelphia outbreak (Fig. 1), collected by Dr. John Neill. By coupling targeted enrichment with high-throughput sequencing,[7] we used trace degraded DNA from this specimen to reconstruct a mid–19th century *V. cholerae* genome and to test the hypothesis that the O1 classical biotype was responsible for the second cholera pandemic.

From the McMaster Ancient DNA Centre (A.M.D., J.M.E., M.K., S.F., K.I.B., H.N.P.), Departments of Anthropology (A.M.D., M.K., K.I.B., H.N.P.), Biology (J.M.E., G.B.G., H.N.P.), and Mathematics and Statistics (D.J.D.E.), and the Michael G. DeGroote Institute for Infectious Disease Research (N.W., D.J.D.E., H.N.P.), McMaster University, Hamilton, ON, and the Dalla Lana School of Public Health, Toronto (D.N.F.) — all in Canada; the Department of Mathematics, Ohio State University, Columbus (J.H.T.); Marie Bashir Institute for Infectious Diseases and Biosecurity Institute, School of Biological Sciences and Sydney Medical School, University of Sydney, Sydney (M.S., E.C.H.); and the College of Physicians of Philadelphia, Mütter Museum, Philadelphia (A.N.D.). Address reprint requests to Dr. Poinar at the Department of Anthropology, McMaster University, 1280 Main St. W., Hamilton, ON L8S 4L9, Canada or at poinarh@mcmaster.ca.

This article was published on January 8, 2014, at NEJM.org.

DOI: 10.1056/NEJMoa1308663

METHODS

We extracted DNA with the use of an organic extraction protocol and converted the extracted DNA into Illumina sequencing libraries[8] in dedicated ancient-DNA facilities. We then enriched the libraries for the O1 classical genome (strain O395; National Center for Biotechnology Information reference sequences NC_009456 and NC_009457) as well as for regions not found in classical strains (e.g., vibrio seventh-pandemic islands I and II [VSP-I and VSP-II]), human mitochondrial genome and amelogenin X and Y genes, and other regions. We mapped the sequencing reads to strain O395 with the use of Burrows–Wheeler Aligner, version 0.5.9rc1 (release 1561),[9] aligned the resulting consensus sequence to 31 full genomes (Table S2 in Supplementary Appendix 1, available with the full text of this article at NEJM.org) using Mauve,[10] called SNPs with a custom PERL script, and visualized features of the sequence in comparison with those of strain O395 with the use of Circos, version 0.36-4[11] (Fig. 2). We used Gblocks[12] to prune insertions, deletions, and potential misalignments, which resulted in a final alignment of 28,591 SNPs that we used in maximum-likelihood[13] and Bayesian[14] phylogenetic analysis.

Our full methods and results are available in Supplementary Appendix 1. Sequences can be found at the Sequence Read Archive (www.ncbi.nlm.nih.gov/sra) under the BioProject accession number SRP029921. Ethics approval for the study was obtained from Hamilton Health Sciences and McMaster University.

Glossary

Ancient DNA: The term used to characterize nucleic acids isolated from an extinct or long-dead organism. Ancient DNA is typically highly degraded and damaged.

Cholera toxin phage (CTX): One of the main virulence factors of *Vibrio cholerae*, located on a genomic island acquired through the CTX prophage. Different pathogenic strains possess different CTX numbers and variants. Cholera toxin itself interferes with the membrane pores of gastrointestinal cells, inducing the water loss that leads to severe diarrhea.

Clade: A phylogenetic group that contains an ancestor and all its descendants.

Genomic islands (GI): Unique genome regions, frequently mobile elements, acquired through lateral gene transfer; these regions can confer specialized functions, such as antibiotic resistance or pathogenicity.

High-throughput sequencing: A suite of modern sequencing technologies that generate millions to billions of unique DNA sequences per run. Also known as next-generation sequencing.

Molecular-clock dating: Dating of divergence times between taxa by estimating the number of substitutions separating them and calibrating with either known times of sampling or fossil ages.

O1 classical: One of the two pathogenic cholera biotypes of the O1 serogroup. Classical cholera is believed to be responsible for the first six global pandemics (until the emergence of El Tor in the 20th century).

O1 El Tor: The second of the two pathogenic cholera biotypes of the O1 serogroup. El Tor strains have been responsible for most of the cholera outbreaks during the past approximately 50 years.

Phylocore genome (PG): The clade that includes all known pandemic cholera strains of serogroups O1 and O139, plus their close relatives. PG-1 and PG-2 are two subclades.

Prophage: A viral genome that has integrated into a bacterial plasmid or chromosomal genome.

Recombination: The process in which novel genetic sequences are formed through the exchange and combination of DNA from two original chromosomes.

RTX: Repeat-in-toxin gene cluster, located next to the CTX locus.

Single-nucleotide polymorphism (SNP): A DNA mutation that is a single-base-pair variant at a position (A, T, G, or C). A SNP is a form of point mutation.

Site saturation: Over evolutionary time, mutations in a DNA sequence can recur at the same site, in many cases reverting to a previous nucleotide. This can reduce, or saturate, the available signal for estimating rates of change or divergence between two sequences, leading to erroneous inferences.

Targeted enrichment: Method to “enrich” for DNA of a certain desired target (e.g., a pathogen genome) by the exposure of synthesized single-stranded “bait” molecules (RNA or DNA) to total prepared single-stranded sample DNA. Baits are either free-floating or fixed to a medium such as a glass slide. Baits and sample are exposed to each other for many hours, which allows for hybridization between any complementary sample and bait DNA sequences. The bait–sample mixture is subsequently washed to remove loosely or poorly bound sample molecules. Any hybridized sample molecules that remain bound to the baits are retrieved and sequenced with the use of high-throughput sequencing.

Vibrio pathogenicity islands (VPI): VPI-1 and VPI-2 are genomic islands (see above) in pathogenic *V. cholerae* strains that are known to function in cholera virulence through the addition of important adhesion and other genes.

Vibrio seventh-pandemic (VSP) islands: VSP-I and VSP-II are genomic islands (see above) found in seventh-pandemic *V. cholerae* strains (clade 7P, in PG-1) but not in other virulent *V. cholerae* strains (such as O1 classical).

RESULTS

HISTORICAL *V. CHOLERAE* GENOME

We reconstructed a draft *V. cholerae* genome (PA1849) at an average unique coverage depth of 15.0×, comprising 94.8% of the O395 reference strain with at least 1.0× coverage and differing by 203 SNPs (see Supplementary Appendix 2). If regions not present in PA1849 (see below) are excluded, then 97.4% of the reference sequence is present at an average coverage depth of 15.4×. Reference strain O395 regions not covered by our sequencing data could represent missing, rearranged, or highly divergent regions in the historical genome itself; could be the result of preservation or procedural biases (e.g., poorly preserved AT-rich regions, biased amplification, and uneven enrichment[7,15]); or both. The *V. cholerae* DNA fragments have typical ancient-DNA damage patterns (Fig. S4 in Supplementary Appendix 1).[16] Overall, coverage correlates strongly with GC content across most regions (Fig. 2, and Fig. S3 in Supplementary Appendix 1).

GENOMIC ISLANDS AND VIRULENCE FACTORS

Strain PA1849 shares the following phylocore genome (PG) genomic islands (GIs) with strain O395: O1, vibrio pathogenicity islands 1 and 2 (VPI-1 and VPI-2), and GI-1 through GI-10 (Table S3 in Supplementary Appendix 1); in addition, PA1849 possesses the PG-2 islands GI-23 (a putative prophage found today only in classical strains O395 and RC27) and GI-24 (a putative prophage with CRISPR [clustered regularly interspaced short palindromic repeat]–associated proteins) (Fig. 2).[4] Strain PA1849 does not have GI-11, GI-14, or GI-21; the absence of these GIs suggests that they were acquired after 1849 by the modern classical strains.

Strain PA1849 contains all known major virulence regions (e.g., VPI-1, VPI-2, and CTX prophage) common to classical *V. cholerae* but does not have nonclassical genomic regions or variants (e.g., VSP-I and VSP-II) (Tables S3 and S4 in Supplementary Appendix 1). The average GC content for these loci is not substantially lower than that in the successfully recovered genomic regions, suggesting that the absence of the loci is unlikely to be an artifact of preservation. VPI-1 has lower-than-average coverage (7.5×, vs. 15.0×), which is probably a result of its relatively low GC content (35%, vs. 46.7% for the entire genome) rather than its absence in the historical genome (Fig. 2).

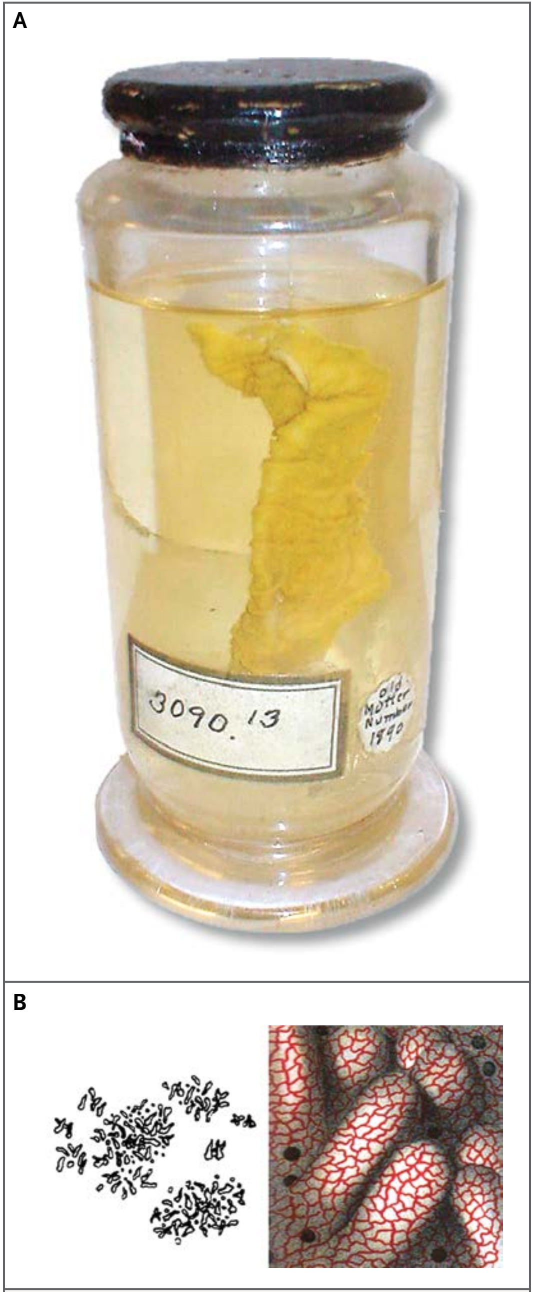


Figure 1. Historical Intestinal Specimen.

Panel A shows specimen 3090.13, a preserved portion of intestine from an 1849 cholera victim. The specimen was collected by Dr. John Neill and colleagues for the purpose of studying the effects of cholera on the intestinal lining. Their 1849 report was presented to the College of Physicians of Philadelphia, and the specimens were displayed, later becoming part of the Mütter Museum collections. Panel B shows examples of histologic and morphologic sketches of intestinal damage that accompanied the report; on the left is a sketch of villi structure, and on the right is a sketch of villi at greater magnification with injected capillaries visible.

Relative to the region in strain O395, VPI-1 in strain PA1849 contains one synonymous SNP (in *tcpA*), and VPI-2 contains four SNPs.

Like strain O395, strain PA1849 contains the classical *ctxB* and *rstR* variants and the expected deletion in the large-chromosome RTX element. It is also likely to possess the same CTX positions as strain O395, because it appears to have identical chromosomal flanking regions, albeit observed at low coverage. However, its CTX prophage configuration, which varies between strains,[4,17] has not been observed in classical

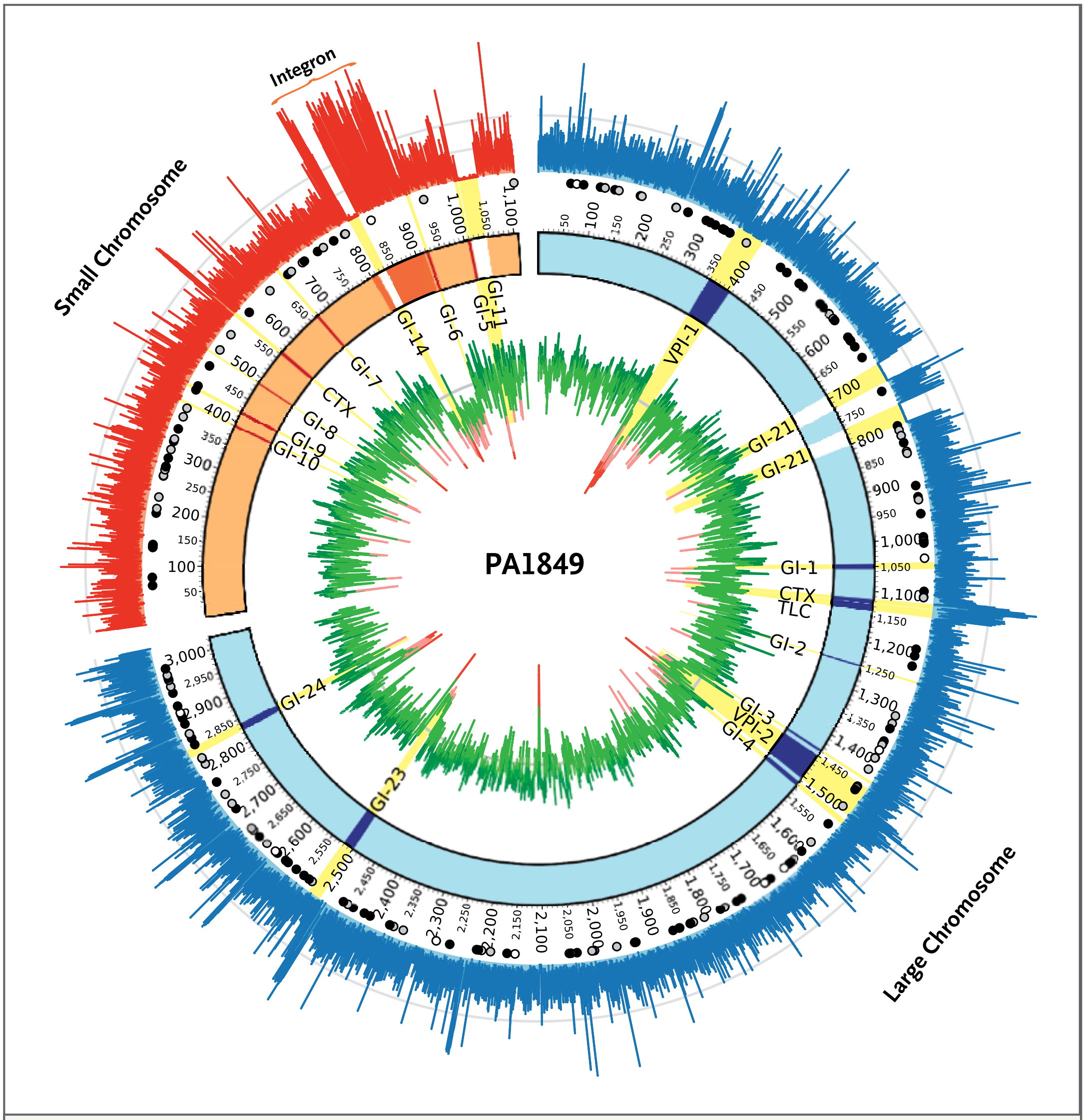


Figure 2. The PA1849 Genome.

Both chromosomes of the PA1849 draft genome, as compared with the O395 reference genome (NC_009456 and NC_009457), are illustrated. The ring closest to the center (in green and red) shows the GC content of O395 across 1000-bp consecutive windows (the gray axis line denotes the genomic average of 47.5%); the next ring shows the chromosomes (with the large chromosome shown in blue and the small chromosome shown in orange) measured in kilobases, with major genomic islands (GIs), vibrio pathogenicity islands (VPIs), cholera toxin phage (CTX), and toxin-linked cryptic (TLC) labeled; the next ring shows single-nucleotide polymorphisms (SNPs) as compared with O395 (black circles denote nonsynonymous, gray synonymous, and white noncoding SNPs); and the outer ring shows unique coverage across 100-bp consecutive windows (the gray axis lines indicate 25× and 50× coverage).

strains to date; read assemblies indicate that there is a tandem CTX repeat span (Fig. S8 and S9 in Supplementary Appendix 1) on one or both chromosomes, with no read assemblies supporting the presence of the truncated CTX prophage repeat that is typical of modern classical strains (Fig. S10 in Supplementary Appendix 1).

HUMAN MITOCHONDRIAL AND NUCLEAR DNA

The complete mitochondrial DNA genome from the patient with cholera was retrieved with a coverage depth of 149.0×, and reads exhibit a typical ancient-DNA damage profile (Fig. S4C in Supplementary Appendix 1).[16] The consensus sequence belongs to haplogroup L3d1b3, found today in sub-Saharan western Africa.[18] Reads matching amelogenin gene X and Y alleles suggest that this patient was male, although coverage across these regions was poor.

ORIGIN AND EVOLUTION OF *V. CHOLERAE*

Our phylogenetic analysis revealed a major division between the PG-1 and PG-2 lineages (Fig. 3). Most El Tor strains cluster in the seventh-pandemic (7P) clade, which also includes strain MO10.[4,19]

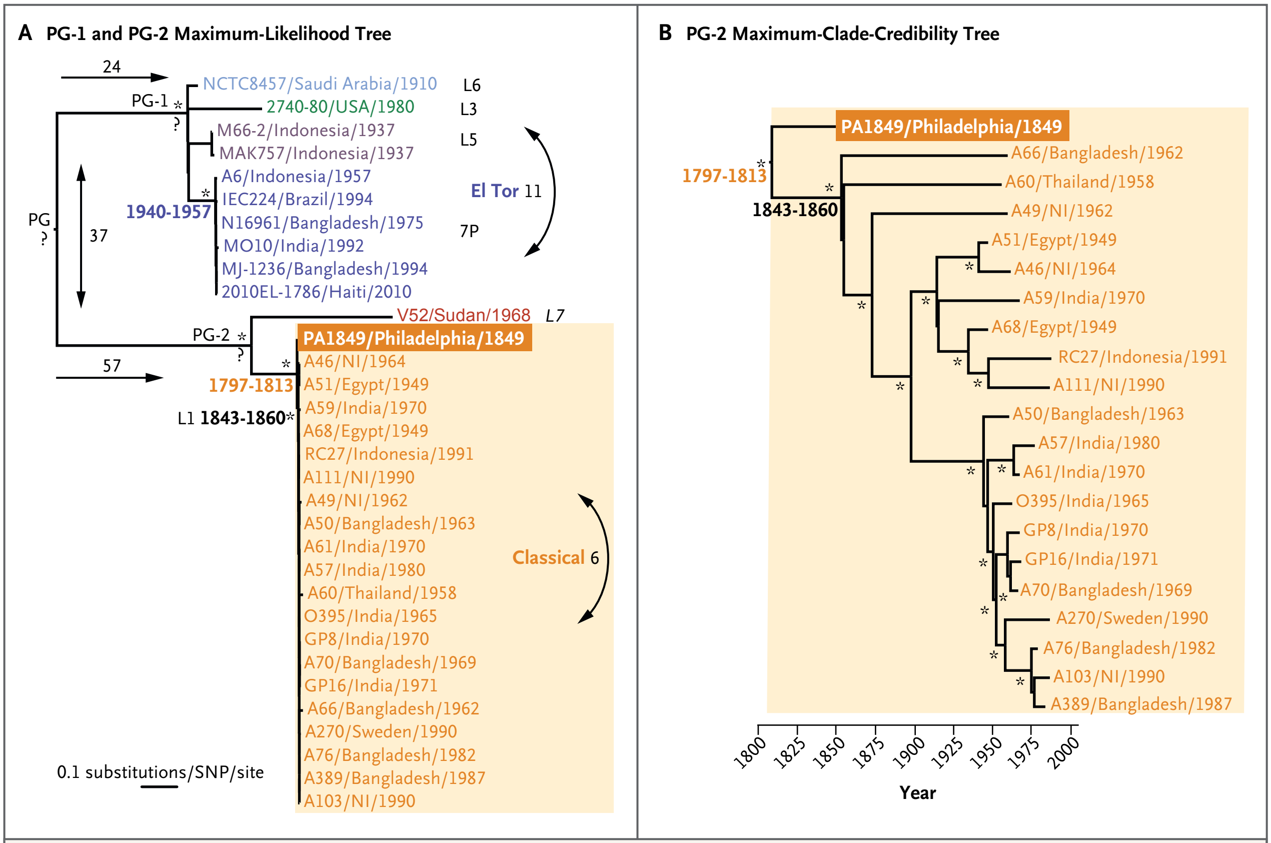


Figure 3. Evolutionary Analysis of *Vibrio cholerae*.

Panel A shows the maximum-likelihood phylogenetic tree of the El Tor and classical (including PA1849) strains (28,591 SNP sites) of *V. cholerae* present within the phylocore genome (PG) groups PG-1 and PG-2. Strain names are expanded to give the place and date of isolation. Branches are color-coded to match the geographic distribution of strains as shown in Figure S6 in Supplementary Appendix 1. Estimates of the divergence times (95% highest posterior density intervals) of key nodes obtained from a separate molecular-clock dating analysis are shown. Question marks indicate key nodes, particularly PG nodes, in which a combination of multiple substitutions and recombinations hindered attempts to accurately estimate divergence times. Branch lengths are scaled to the number of nucleotide substitutions per SNP site, and bootstrap support values greater than 90% at key nodes are indicated with an asterisk. The numbers of recombination events within (curved double-headed arrows) and between (vertical double-headed arrow) the PG-1 and PG-2 groups, as well as into PG-1 and PG-2 from unknown parents (horizontal arrows), are also shown. Panel B shows the maximum-clade-credibility tree for the evolution of the PG-2 classical strains (excluding L7) in inferred chronologic time. All tip times are set to the time of sampling, as shown on the x axis. Posterior probability values greater than 0.90 are indicated with an asterisk. 7P denotes the seventh-pandemic clade.

Representatives of clades L6 (strain NCTC 8457), L3 (strain 2740-80), and L5 (strains M66-2 and MAK757), together with the 7P clade, make up the PG-1 clade, whereas the PG-2 clade comprises L7 (strain V52), L1 (classical strain), and strain PA1849. Strain PA1849 sits several SNPs away from the L1 clade node, with strong bootstrap support.

Our initial attempts to date the evolutionary history of *V. cholerae* were hindered by an inability to estimate evolutionary rates from tip dates alone, even when PG-1 and PG-2 were analyzed separately. This is probably the result of a combination of site saturation and the extensive recombination that is typical of *V. cholerae*.[4,19,20] We confidently detected at least 37 recombination events between PG-1 and PG-2 (Fig. 3), as well as a number of intragroup recombination events and events that brought genetic diversity into PG-1 and PG-2 from unknown parental lineages. Such a recombination frequency makes it difficult to determine whether specific recombination events explain the recent predominance of El Tor strains.

To overcome these limitations, we imposed a strict molecular clock (1.3×10^{-3} nucleotide substitutions per SNP site per year) based on a reanalysis of a large El Tor data set.[19] Assuming this rate, we estimate that the El Tor 7P strains emerged between 1940 and 1957 (95% highest posterior density), in agreement with previous estimates.[19] Similarly, we estimate that the ancestor of the classical strains originated between 1843 and 1860, with divergence of the lineage leading to strain PA1849 occurring between 1797 and 1813, close to the time of the first recognized cholera pandemic, in 1817.[21] The combined topology and temporal estimations suggest that the first five cholera pandemics were caused by *V. cholerae* possessing a common core genome, each representing a clonal reemergence with few genome-scale mutational differences.

DISCUSSION

The PA1849 genome has a number of unique structural features but differs from strain O395 by only a few hundred SNPs across the entire 4-Mb genome. This suggests that modern *V. cholerae* evolution has been subjected to substantial selective constraint since the mid-19th century, similar to that of other pathogens that exhibit long-term core genome sequence conservation over a period of centuries (e.g., *Yersinia pestis*[7] and *Mycobacterium leprae*[22]). One of the striking features of the PA1849 historical genome is the tandem CTX configuration; this could indicate that it was capable of producing infectious CTX virions,[23] which potentially conferred greater pathogenic capacity.[24] Therefore, the suggestion that the absence of CTX virion production in classical strains may have contributed to their replacement by El Tor[24] may be unfounded. However, because there is a *V. cholerae* strain (B33) with tandem CTX repeats that does not produce replicating virions,[25] the functional implications of this structure in strain PA1849 cannot be confirmed without experimental expression in a model vibrio strain.

Previous attempts to date the origin of pandemic *V. cholerae* have returned very different results (e.g., approximately 130 to 50,000 years ago for the classical–El Tor split[26]). If a constant evolutionary rate of 1.3×10^{-3} substitutions per SNP site per year were applied across the entire phylogeny, then the common ancestor of all pathogenic *V. cholerae* (PG, potentially the ancestral strain first adapted to humans) would date to only 430 to 440 years ago. However, a combination of site saturation and recombination means that this date is an underestimate, and the date of PG is more likely to be on a time scale of millennia, predating all historically recognized pandemics and arguing against a postmedieval origin of pathogenic *V. cholerae*.[26] Our analysis therefore suggests that the PG-1 and PG-2 lineages cocirculated in humans and water sources for many centuries and potentially thousands of years before the 19th century pandemics, a finding compatible with the theory that cholera is a disease of the "first epidemiological transition," during which sedentary agriculture (beginning approximately 10,000 years ago) opened new disease niches.[27]

Collections of historical pathological specimens are invaluable resources for reconstructing pathogen evolution, yet the study of these collections remains a sensitive topic, because it was common for the bodies of marginalized minorities and the poor to be retained for medical research without consent.[28] We hope that by highlighting the intrinsic scientific, historical, and social value of these underappreciated collections, we can help to recognize and protect them in perpetuity.

Supported by Canada Research Chairs from the Natural Sciences and Engineering Research Council of Canada (to Drs. Poinar and Golding), a grant from the Social Sciences and Humanities Research Council of Canada (to Dr. Poinar), a National Health and Medical Research Council Australia Fellowship (to Dr. Holmes), and an Ontario Graduate Scholarship (to Ms. Devault).

Disclosure forms provided by the authors are available with the full text of this article at NEJM.org.

We thank Dr. D. Ann Herring, Dr. D. Poinar, Dr. C. Yates, Dr. G. Wright, and current and former members of the McMaster Ancient DNA Centre, McMaster University, Hamilton, ON, Canada.

REFERENCES

1. Chin CS, Sorenson J, Harris JB, et al. The origin of the Haitian cholera outbreak strain. N Engl J Med 2011;364:33-42.

2. Ali M, Lopez AL, You YA, et al. The global burden of cholera. Bull World Health Organ 2012;90:209A-218A.

3. Faruque SM, Chowdhury N, Kamruzzaman M, et al. Genetic diversity and virulence potential of environmental Vibrio cholerae population in a cholera-endemic area. Proc Natl Acad Sci U S A 2004;101:2123-8.

4. Chun J, Grim CJ, Hasan NA, et al. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. Proc Natl Acad Sci U S A 2009;106:15442-7.

5. Rosenberg CE. The cholera years — the United States in 1832, 1849, and 1866. Chicago: University of Chicago Press, 1987.

6. Harris JB, LaRocque RC, Qadri F, Ryan ET, Calderwood SB. Cholera. Lancet 2012;379:2466-76.

7. Bos KI, Schuenemann VJ, Golding GB, et al. A draft genome of Yersinia pestis from victims of the Black Death. Nature 2011;478:506-10. [Erratum, Nature 2011;480:278.]

8. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protoc 2010;2010(6):prot5448.

9. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.

10. Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One 2010;5(6):e11147.

11. Krzywinski M, Schein J, Birol I, et al. Circos: an information aesthetic for comparative genomics. Genome Res 2009;19:1639-45.

12. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56:564-77.

13. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Syst Biol 2010;59:307-21.

14. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. BMC Evol Biol 2007;7:214.

15. Briggs AW, Stenzel U, Johnson PLF, et al. Patterns of damage in genomic DNA sequences from a Neandertal. Proc Natl Acad Sci U S A 2007;104:14616-21.

16. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 2011;27:2153-5.

17. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 1996;272:1910-4.

18. Barbieri C, Whitten M, Beyer K, Schreiber H, Li MK, Pakendorf B. Contrasting maternal and paternal histories in the linguistic context of Burkina Faso. Mol Biol Evol 2012;29:1213-23.

19. Mutreja A, Kim DW, Thomson NR, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 2011;477:462-5.

20. Kaper JB, Morris JG Jr, Levine MM. Cholera. Clin Microbiol Rev 1995;8:48-86. [Erratum, Clin Microbiol Rev 1995;8:316.]

21. Barua D. History of cholera. In: Barua D, Greenough III WB, eds. Cholera. New York: Plenum Medical, 1992:1-36.

22. Schuenemann VJ, Singh P, Mendum TA, et al. Genome-wide comparison of medieval and modern Mycobacterium leprae. Science 2013;341:179-83.

23. Davis BM, Moyer KE, Boyd EF, Waldor MK. CTX prophages in classical biotype Vibrio cholerae: functional phage genes but dysfunctional phage genomes. J Bacteriol 2000;182:6992-8.

24. Davis BM, Waldor MK. CTXphi contains a hybrid genome derived from tandemly integrated elements. Proc Natl Acad Sci U S A 2000;97:8572-7.

25. Faruque SM, Tam VC, Chowdhury N, et al. Genomic analysis of the Mozambique strain of Vibrio cholerae O1 reveals the origin of El Tor strains carrying classical CTX prophage. Proc Natl Acad Sci U S A 2007;104:5151-6.

26. Feng L, Reeves PR, Lan R, et al. A recalibrated molecular clock and independent origins for the cholera pandemic clones. PLoS One 2008;3(12):e4053.

27. Armelagos GJ, Brown PJ, Turner B. Evolutionary, historical and political economic perspectives on health and disease. Soc Sci Med 2005;61:755-65.

28. Harrington JM, Blakely RL. Rich man, poor man, beggar man, thief: the selectivity exercised by graverobbers at the Medical College of Georgia, 1837-1887. In: Saunders SR, Herring A, eds. Grave reflections: portraying the past through cemetery studies. Toronto: Canadian Scholars' Press Toronto, 1995:153-78.

Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Devault AM, Golding GB, Waglechner N, et al. Second-pandemic strain of *Vibrio cholerae* from the Philadelphia cholera outbreak of 1849. N Engl J Med. DOI: 10.1056/NEJMoa1308663

"Second-Pandemic Strain of *Vibrio cholerae* from the Philadelphia Cholera Outbreak of 1849"

SUPPLEMENTARY APPENDIX

TABLE OF CONTENTS

AUTHORS

Alison M. Devault, M.A., G. Brian Golding, Ph.D., Nicholas Waglechner, M.Sc., Jacob M. Enk, M.Sc., Melanie Kuch, M.Sc., Joseph H. Tien, Ph.D., Mang Shi, Ph.D., David N. Fisman, M.D., M.P.H., Anna N. Dhody, M.F.S., Stephen Forrest, M.Sc., Kirsten I. Bos, Ph.D., David J.D. Earn, Ph.D., Edward C. Holmes, Ph.D., and Hendrik N. Poinar, Ph.D.

I. SUPPLEMENTARY METHODS

A. Cholera Victim Specimen

The intestinal specimen (museum catalog number 3090.13) from a putative 1849 cholera victim examined in this paper is part of the collections of the Mütter Museum (College of Physicians; Philadelphia, PA, USA). The original description associated with the sample is as follows: "*Intestine. Cholera. Yellow. Presented by Dr. John Neill.*" No date is directly associated with the specimen, and no museum records appear to directly reference this specimen. However, this set of specimens (of which at least 7 remain) that purport to be from cholera victims is almost certainly part of the exhibited collection presented by Neill in 1849 to the College of Physicians (see quote below). We have been unable to locate the dry preparations, microscope preparations, or drawings from this set. Also it is unclear which (if any) of the wet preserved specimens belonged to the control set of specimens (from the victims of pleurisy).

> "In 1849, the Asiatic cholera visited us for the second time, and Neill was physician to the Southeast Cholera Hospital, which was established in the old Friends' Meeting House, in Pine Street, above Front. During his service here he made the minute injections upon which was based the Report made to this College, and published in the Transactions, as we shall state more fully hereafter. … At the meeting of Dec. 4, 1849, the Committee on Cholera, of which Professor Samuel Jackson was Chairman, and Drs. Neill, H. H. Smith and Pepper were members, presented a report upon the injections of the intestines of patients dying of cholera, which had been made by Dr. Neill, as has already been stated. The Committee reported that "the admirable manner in which he performed this duty can be judged of by the beautiful preparations now upon the table, which he has presented to the College for its Museum." The injections were made with turpentine colored with vermilion, having failed when size and Canada balsam had been used. This had led, at first, to the supposition that the capillaries of the intestinal tract were destroyed by the disease; but the method devised by the investigator showed the perfect integrity of those vessels. A portion of healthy intestine, taken from the body of a patient who had died of pleurisy, was injected with the same fluid as a standard of comparison. These preparations speak for themselves; they form a groundwork for any future investigation, and those interested may find them in the Museum of this building. The whole series consisted of eleven mounted wet, and twelve dry preparations, four fitted for the microscope, and seven drawings. An interesting discussion ensued upon the presentation of these specimens to the College." [1]

B. Sampling

Subsampling of the 3090.13 intestinal tissue specimen was performed at the Mütter Museum's curation room by AMD and JT. The workspace was thoroughly bleached prior to subsampling. The lid of the sealed jar was removed by cutting along the circumference of the lid using a thin dremel wheel (diamond-tip) and scalpel blade. The tissue specimen was removed from the solution with tweezers and placed into a clean weighing boat. A small section (~2x2 cm) was transferred to a sterile 50-mL falcon tube by slicing off the end of the intestinal specimen with a scalpel blade. The specimen was returned to the jar and the jar opening was temporarily sealed with plastic tape. The curatory staff restored the jar's seal after subsampling was complete. A secondary subsample was also taken at the time of sampling to retain in frozen storage at the Mütter Museum for future work.

C. DNA Extraction

All laboratory methods were performed in the facilities of the McMaster Ancient DNA Centre (McMaster University, Hamilton ON, Canada), which include physically separated "ancient" clean room (extraction, library preparation, and PCR set-up) and "modern" laboratory (amplification, enrichment, and re-amplification) facilities. Subsampled 3090.13 tissue specimen was stored at -20°C upon arrival. An approximately 92 mg tissue sample was subjected to DNA extraction using a modified tissue extraction protocol[2], alongside extraction blanks (1 per set of extractions). Tissue sample was divided into small fragments using a scalpel blade and incubated with 800µl of tissue digestion buffer (pH 7.5; 1% SDS, 50mM DTT, 25 mM Tris-HCl pH 8.0, 25 mM sodium citrate, 5 mM $CaCl_2$, 2.5 mM EDTA pH 8.0, 10 mM PTB, and water) for 95°C for 5 min (1000 rpm). We added 80 µl of proteinase K (20 mg/mL, activated), and the contents were incubated with rotation for 50°C for 24 hours. Digested material was subjected to organic extraction with 2x 800 µl phenol-chloroform (spun for 10 min at 14,000 rpm and aqueous phase retained) and 2x 800 µl chloroform. The organic extraction products were subjected to column ultrafiltration with 10 KDA cut-off (via 10K Amicon Ultra-0.5mL; Millipore, MA). Column was primed with 300 µl of 0.1x TE and spun to minimum retention. Extracted material was applied stepwise to column and spun to minimum retention. Three washes of 300 µl 0.1xTE were applied to the column and spun to minimum retention. Filters were inverted into clean microcentrifuge tubes and spun at 1000 RCF for 5 minutes to capture extracted DNA solution (if necessary, volume was increased to reach a final volume of 50 µl 0.1x TE for the extracted DNA solution).

D. Library Preparation and Indexing

Preparation of adapted libraries capable of sequencing on the Illumina platform was performed as given in reference 3, with some modifications as given in reference 4. 450 µl of a 1 in 10 dilution (in 0.1x TE) of extracted 3090.13 DNA solution was used as template in multiple library preparation reactions (50 µl template DNA in each of 9 libraries). 45 µl of extraction blank library plus 5 µl 0.1x TE was also used as template in an additional library preparation reaction. Two library preparation blanks (0.1xTE; water) were also included. Library preparation enzymes were purchased from New England Biolabs (NEB) and used in these concentrations during library preparation: T4 polynucleotide kinase, 0.5 U/µl; T4 DNA polymerase, 0.1 U/µl; T4 DNA ligase, 0.125 U/µl. MinElute (Qiagen) PCR product purification was substituted for SPRI bead clean-up between library preparation steps; instead of a final purification after the fill-in stage, a 20 minute 80°C heat inactivation was performed, and the final product was used as template in indexing PCRs directly.[4] After library preparation, the 9 3090.13 libraries were pooled to a final library volume of 360 µl. The three blanks had final volumes of 40 µl.

Post-library quantitation. Primers *IS7_short_amp.P5* and *IS8_short_amp.P7* [3] were used in a quantitative PCR assay alongside a quantitative library standard (49bp template insert) to quantify "copies/µl" (c/µl) of the libraries: 2.2E07 3090.13, 4.3E03 extraction blank, 6.3E03 0.1xTE blank, 8.7E03 H_2O blank. The qPCR conditions were 200 nM each primer, 1X PCR Buffer II (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 250 µM dNTPs, 0.15 mg/mL bovine serum albumin, 0.05 U/ µl Amplitaq Gold polymerase (Applied Biosystems), 0.167X SYBRgreen fluorescent dye (Invitrogen), and 1 µl template DNA (1 in 100 or 1:1K dilution) in a 10 µl final qPCR volume. A 95°C initial denaturation (4 min) followed by 50 cycles of 95°C (30s), 62°C (30s), and 72°C (30s) were performed.

Indexing. Double indexing [4] PCR reactions were performed as given in reference 3, using AmpliTaq Gold polymerase instead of Phusion Hot Start High-Fidelity DNA polymerase: 14x 3090.13, 2x extraction blank, 2x each library blank, and 2x PCR blanks. PCR conditions were 200 nM each indexing primer, 1X PCR Buffer II, 2.5

mM $MgCl_2$, 250 µM dNTPs, 0.15 mg/mL bovine serum albumin, 0.05 U/ µl Amplitaq Gold polymerase, and 10 µl DNA template, in a 100 µl final PCR volume. A 95°C initial denaturation (4 min) followed by 10 cycles of 95°C (30s), 60°C (30s), and 72°C (30s) were performed. Indexed reactions were pooled by sample/blank and every 200 µl was purified over MinElute to a final volume of 140 µl 3090.13 and 20 µl each blank.

E. *V. cholerae* DNA qPCR Assessment

A qPCR assay specific to the cholera genome was designed to amplify a 58bp fragment of the *ompW* gene (forward primer Vchol_ompW_507F GCTCAATGATAGCTGGTTCCTCAAC; reverse primer Vchol_ompW_564R CGTTGTTTCAATATTGGCATACCACAC). A 580bp fragment of the *ompW* gene was amplified from the 3090.13 sample (results not shown) and cloned to serve as a standard control of known copy number for qPCR. The 58bp *ompW* assay was optimized for sensitivity down to <10 c/µl under the following PCR conditions: 200 nM each primer, 1X PCR Buffer II (Applied Biosystems), 2.5 mM $MgCl_2$ (Applied Biosystems), 250 µM dNTPs, 0.05 U/µl Amplitaq Gold polymerase (Applied Biosystems), 0.5X EVAgreen fluorescent dye (Biotium), and 1 µl template DNA in a 10 µl final qPCR volume, with cycling parameters of 95°C initial denaturation (4 min) followed by 50 cycles of 95°C (30s), 65°C (30s), and 72°C (30s). Amplifications using these conditions were performed (in replicate) on the following templates at 1 in 100 (0.01x) dilutions (in 0.1xTE): libraries (3090.13, extraction blank, 2x library preparation blanks), indexed libraries (3090.13, extraction blank, 2x library preparation blanks), and re-amplified libraries (extraction blank, 2x library preparation blanks, and indexing blank). Standards of known copy number (10K, 1000, 100, 10) and PCR blanks were also included in replicate. Positive amplifications of the 58bp product were only detected in the 3090.13 templates, at the following average calculated c/µl: 9.22E03 library, 9.76E03 indexed library (see **Figure S1**). All tested blanks (as indicated above) were free of the 58bp cholera specific product.

F. Array Design

DNA enrichment was performed using a custom one-million feature Agilent SureSelect DNA Capture Array (Agilent Technologies, Inc., Santa Clara, CA). All probes were designed using software described elsewhere.[5] Efforts were made to identify and mask repetitive elements based on 15-mer frequency counts in the design template, where all probes carrying an average 15-mer frequency of 100 or more were eliminated.[6] In addition, probes containing a 15-mer sequence represented in either of the Illumina adapter sequences (P7, CAAGCAGAAGACGGCATACGAGATGTGACTGGAGTTCAGACGTGT and P5, AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTT) were removed to discourage non-specific binding of adapters. Probes with base pair sequences identical to those already generated were subsequently removed. Both chromosomes of the *Vibrio cholerae* O395 strain (O1 classical serotype) were included on the array at 5bp tiling, with increased tiling density (2bp offset) across several specific regions of interest (see **Table S1**). Other virulence regions associated with other modern pathogenic cholera strains such as El Tor and O139 (but NOT expected in O1 classical) were also included on the array at 2bp tiling, such as VSP-I & VSP-II, the El Tor *tcpA* variant, and the O139 antigen region. Other probes were also included on the array at 4bp tiling density to enhance the recovered information on the historical cholera victim: human mitochondrial genome and human X and Y amelogenin genes. The array also included 13,729 additional non-project probes (results not discussed here). The total number of probes included on the array was 923,153.

G. Array Enrichment

Pre-enrichment amplification. Prior to enrichment, additional indexing amplification was performed in 32 reactions (2.5 µl straight template DNA in 50 µl total reaction volume) of the 3090.13 library (see above) using 400 nM each indexing primer and 15 cycles to reach a final, post-MinElute purified total DNA concentration of 187 ng/µl.

Enrichment. Array enrichment was performed as in reference 7 and the Agilent SureSelect DNA Capture Array protocol (v1.0)[8]. 18.7 total µg of amplified 3090.13 DNA (100 µl) was combined with 18 µl nuclease-free water, 40 µl of 8 blocking oligos (2 µM), 50 µl Cot-1 DNA (0.1 mg/mL), 52 µl 1X Agilent Blocking Agent, and 260 µl 1X Agilent Hi-RPM Hybridization Buffer. This mixture was distributed into low-bind PCR tubes and incubated at 95°C (3 min) and 37°C (30 min), pooled, transferred to the enrichment array slide sandwich in the hybridization chamber, and rotated in the dark at 10 rpm for 65h at 65°C. The enriched fraction was recovered using the protocol of Hodges et al (2009) and Agilent, with a final recovered eluate volume of 462.5 µl. (We also recovered the post-enrichment "non-enriched" fraction by additionally performing the syringe protocol prior to slide washing).

Post-enrichment amplification. 450 µl of post-enriched eluate alongside PCR blanks was immediately amplified in 18x qPCR reactions using the re-amplification primers *IS5_reamp.P5* and *IS6_reamp.P7* [3] according to the following protocol: 400 nM each primer, 1X PCR Buffer II, 2.5 mM $MgCl_2$, 250 µM dNTPs, 0.05 U/µl Amplitaq Gold polymerase, 0.167X SYBRgreen fluorescent dye (Invitrogen), and 25 µl DNA template, in a 50 µl final PCR volume. A 95°C initial denaturation (4 min) followed by 20 cycles of 95°C (30s), 65°C (30s), and 72°C (30s) were performed. Reactions were pooled and purified over MinElute to a final volume of 135 µl (3090.13). qPCR quantitation of pre- and post-amplified eluate indicated a 7.7E04 fold increase due to post-enrichment amplification (1.0E04 c/µl vs. 7.7E08 calibrated c/µl, respectively). (The additional recovered non-enriched, non-sequenced fraction, as noted above, was 2.1E11 c/µl.)

H. Sequencing

The final 3090.13 post-enriched library sent for sequencing was 6.8 ng/µl (via Nanodrop; Thermo Scientific) and 14.5 nM (via Bioanalyzer; Agilent Technologies, Inc). Sequencing was performed across 5 lanes of Illumina GAIIx system by the Donnelly Sequencing Centre (University of Toronto). 72bp single read chemistry was used, but the total read length was extended to 80bp by the addition of 8bp worth of sequencing reagents to the sequencing read length (rather than being used towards an indexing read). Each lane yielded 34,019,321 to 34,821,855 total reads passing filter, in total yielding 172,138,676 raw reads passing filter. We have deposited all raw sequencing reads at the NCBI Sequence Read Archive (http://www.ncbi.nlm.nih.gov/sra) under the BioProject accession number SRP029921.

I. Shotgun Sequencing

For the purposes of comparing the overall difference between percent of *V. cholerae* DNA in the non-enriched vs. enriched state, we also subjected an additional aliquot of indexed library (from section D above) to re-amplification, and sequenced this non-enriched library on the Illumina platform. (It is noted that the number of cycles experienced by the enriched and non-enriched libraries are unequal, so the comparison between pre- and post-enriched percent *V. cholerae* DNA, as discussed in the main manuscript, is not ideal; however, this does not influence any of the main conclusions of this study.)

Pre-sequencing amplification. Prior to sequencing, additional indexing amplification was performed in 8 reactions (5 µl 0.1x diluted template DNA in 50 µl total reaction volume) of 3090.13 indexed library (see above) using 400nM each indexing primer and 11 cycles.

Sequencing. The purified 3090.13 library was pooled with an additional (un-related) sample in equimolar ratio on one lane of Illumina HiSeq 1000. Sequencing was performed by the Farncombe Family Digestive Health Research Institute (McMaster University). 100bp paired-end read chemistry was used, with one indexing read. The lane yielded 263,870,537 total raw reads passing filter, 141,039,627 of which belonged to sample 3090.13.

J. Authorship Contributions

Grants awarded to HNP funded the study. AMD, JME, MK, JHT, DNF, AND, DJDE, and HNP designed the study. AMD, JME, MK, JHT, and HNP gathered the data. AMD, GBG, NW, JME, MK, MS, DNF, SF, KIB, ECH, and HNP analyzed the data. All authors vouch for the data and analysis. AMD, JME, ECH, and HNP wrote the first draft and co-wrote the additional drafts of paper with the assistance of all other co-authors. AMD and HNP decided to publish the paper with the approval of all other co-authors.

II. SUPPLEMENTARY ANALYSIS

A. Draft genome sequence

i. *Trimming.* Trimming was done using cutadapt (version 1.0)[9] to eliminate adapters and to eliminate bases with a PHRED quality cutoff 20, a minimum overlap 3bp between read and adapter, and a minimum length after trimming of 20bp. The qualities of the reads were double-checked using FASTQC.

ii. *Assemblies*. Reads were mapped to *Vibrio cholerae* O395 (NC_009456, NC_009457) using BWA v.0.5.9rc1 (r1561)[10] at default settings. The resulting sam file was edited to retain entries with a minimum length of 20bp, a minimum of 98% of the read length matched perfectly, and a map quality of 20. These were then converted to bam files via samtools[11] and sorted, PCR duplicates were removed using samtools rmdup (-s), and samtools mpileup was used to generate a bcf file. The 'loose' assembly was done with default base quality and map quality (13 and 0, respectively) in mpileup. 'Strict' mappings were treated the same way but with map quality of 20. Fragment length distributions can be found in **Figures S2a & S2b**. Average coverage vs. %GC for the PA1849 genome is plotted in **Figure S3**.

iii. *Ancient DNA Damage Assessment.* We subjected the post-trimmed reads to analysis using the MapDamage (v.0.3.6) program[12] (**Figures S4a & S4b**). Patterns of damage (C→T or G→A changes at the 5’ and 3’ fragment ends) typical to ancient DNA were seen, despite the differences in tissue, age, preservation method, etc. between this sample and other “typical” ancient DNA specimens.

iv. *SNP Calling.* SNPs (n=203) were called from BCF files generated by samtools, using a minimum coverage of 5x, 90% variant frequency, a minimum quality PHRED value of 5 and a mapping quality of ≥ 20 (see above). We also generated alternate SNP datasets that show what SNPs would have been called by restricting base quality values to either minimum 10 (n=192) or 20 (n=147). All SNP calls can be found in **Supplementary Appendix 2**. If by chance some reads of sufficient coverage mapped to locations that were within broader regions that were determined to be actually absent in PA1849 (e.g., GI-21), or within regions that were likely to be conserved within other species or could have similar repetitive regions within the cholera genome (e.g. 16S rRNA or tRNA loci), these were manually removed from the list of final SNPs. Several SNPs were also eliminated that relied on the ends of reads, as these were likely to contain damaged sites due to the nature of "ancient DNA" (see above).

v. *Evolutionary Analysis.* A complete genome sequence alignment of 32 strains of *V. cholerae*, including PA1849 (**Table S2**; map in **Figure S5**), was generated using MAUVE assuming default parameters. (Consensus sequences from any SRA database readsets were first generated via assembly to O395.) A SNP alignment was then generated from this complete genome alignment (see "SNP calling"). Because some regions of the complete genome, and hence SNP, alignment were uncertain, reflecting the occurrence of multiple indels, we used Gblocks[13] to purge all ambiguous regions, including all indel sites. This resulted in a final SNP alignment of 28,591 nt that could be utilized for evolutionary analysis. Although we also undertook phylogenetic analyses with two environmental strains – 12129/Australia/1985 and LMA3984-4/Brazil/2007 – they were excluded from the final data set because of their highly divergent nature and hence likely extensive site saturation compared to PG1 and PG2 strains.

We first inferred a phylogenetic tree using the Maximum Likelihood (ML) approach available in the PhyML package.[14] This analysis employed the GTR model of nucleotide substitution and a gamma distribution of among-site rate variation with four rate categories (Γ_4) (i.e. the GTR+Γ_4 nucleotide substitution model) and utilizing subtree pruning and regrafting (SPR) branch-swapping until the highest likelihood was obtained (log likelihood = - 153884.6). The final parameter values for the nucleotide substitution model were: GTR = A-C 0.962, A-G 5.548, A-T 0.749, C-G 0.609, C-T 5.591, G-T 1.000; I = 0; Γ_4 = infinity. To determine the robustness of individual nodes on this phylogeny, we performed a bootstrap resampling analysis, employing 1000 pseudo-replicate ML trees estimated using the same procedure as described above, optimizing parameters in each run. This tree is shown in Figure 3 in the main text.

Next, we attempted to infer a time-scale for the evolutionary history of *V. cholerae* using the Bayesian Markov Chain Monte Carlo (MCMC) approach available in the BEAST package (version 1.7.5)[15] and employing the sampling dates (i.e. years) of each sequence, which ranged from 1849-2010. Multiple runs were undertaken using both strict and relaxed (uncorrelated lognormal) molecular clocks, different nucleotide substitution models (i.e. HKY85 and GTR), and different tree priors (i.e. constant population size and Bayesian skyline). All runs were performed multiple times, with chain lengths varying from 100,000,000 to 1,000,000,000. However, the use of these time-informed (i.e. tip-dated) data to estimate rates of nucleotide substitution and times to common ancestry resulted in unstable values in all cases (i.e. poor ESS values, strongly autocorrelated traces), indicative of a lack of temporal signal in the data. A lack of temporal signal was also apparent in a regression of sampling time (year) against root-to-tip genetic distances on the ML tree performing using the Path-O-Gen program kindly provided by Andrew Rambaut, University of Edinburgh (http://tree.bio.ed.ac.uk/software/pathogen/); the correlation coefficient was very low (0.136), confirming a lack of temporal signal in the data.

The lack of temporal signal in the *V. cholerae* SNP data necessitated that we set (i.e. 'fixed'), rather than estimate, a nucleotide substitution rate. To this end we conducted a BEAST analysis of SNP data from 122 El Tor strains (sequence alignment of 1757 bp; data kindly provided by Ankur Mutreja, Wellcome Trust Sanger Institute, UK) sampled between 1937-2010. These data exhibited a strong temporal signal in a root-to-tip regression (correlation coefficient = 0.863) suggesting that they provide a robust estimate of substitution rate, at least over the time span of sampling. Accordingly, a BEAST analysis (strict molecular clock) of these data produced an evolutionary rate of 1.3 x 10^{-3} nucleotide substitutions per SNP site, per year (subs/SNP site/year) (and all parameters converged after chains of 100,000,000 and with good mixing). This rate was then to estimate the time-scale of evolution of (i) the 6 El Tor strains, and (ii) the 21 classical strains, including PA1849, in our 32 sequence data set (i.e. we assume that the El Tor and classical strains have evolved at the same rate). However, because of a combination of site saturation on deep branches (note the scale bar in Figure 3) and recombination (see below) we were unable to obtain robust estimates for the age of deep nodes on the *V. cholerae* phylogeny, including the common ancestor of PG1 and PG2 (i.e. node PG).

vi. *Recombination Analysis.* To identify potential recombinant regions and their evolutionary origin, we utilized seven recombination detection methods implemented in the RDP4 package; RDP, GENECONV, BOOTSCAN, MAXCHI, CHIMAERA, SISCAN, and 3SEQ.[16] Recombination events significantly detected by more than two methods (bonferroni corrected *P* values < 0.05) with default parameters were counted as *bona fide*, and those sharing the same recombination pattern were merged and treated as a single recombination event. We then summarized the occurrences of recombination between lineages, including between the classical and El Tor strains, to produce a final count of the number recombination events across the data set as whole. Note that in some cases the recombinant parent (i.e. donor sequence) was unidentified, and high levels of sequence similarity made it difficult to accurately infer the number of recombination events within the classical and El Tor groups.

To determine whether recombination has compromised our attempts to reconstruct the phylogenetic history and time-scale of *V. cholerae* evolution we repeated the analyses described in section (v) on a data set in which the recombination events defined using RDP4 described above were coded as gaps (-); this allowed us to retain a sequence alignment of the same length as the original (28,591 nt; thereby maximizing phylogenetic resolution) but with all recombinant regions removed. This analysis resulted in similar phylogenetic trees to that obtained with recombinant regions included in the alignment, particularly with respect to the nodes of interest; ML and maximum parsimony trees are shown in **Figures S6 and S7**, respectively. Similarly, to assess the affect of recombination on our molecular clock dating we repeated the BEAST analysis above (with an evolutionary rate set to 1.3 x 10^{-3} subs/SNP site/year) on the recombination-free data set. Strikingly similar times to the Most Recent Common Ancestor (tMRCA) of the whole tree (i.e. node PG) were obtained both with (430-440 years ago) and without (468-487 years ago) recombinant regions, suggesting that site saturation is of more evolutionary importance on these data than recombination. Finally, to obtain an approximate estimate of the number of recombinant sites in our SNP alignment we estimated a number of parsimony-based statistics on the ML tree (with all recombinant regions included). These metrics are as follows: (i) Tree length = 30954; (ii) Consistency index (CI) = 0.9253; (iii) Homoplasy index (HI) = 0.0747; (iv) CI excluding uninformative characters = 0.8744; (v) HI excluding uninformative characters = 0.1256; (vi) Retention index (RI) = 0.9811; and (vii) Rescaled consistency index (RC) = 0.9079. Hence, these results indicate that there is relatively little recombination in these data relative to divergent evolution (i.e. 92.5% of SNPs are consistent with the ML tree and 7.5% of SNPs are homoplasic, perhaps due to recombination).

B. Other Loci Included on the Array: Non-classical *V. cholerae* and Human

i. *Trimming.* The 172,138,676 raw sequenced reads were trimmed to remove any remaining adaptor sequence using *cutadapt* (v.1)[9] with the settings: error rate (0.16), minimum overlap (1).

ii. *Assemblies.* Assemblies to the reference sequences included on the array were done in Geneious (v.5.6.5; Biomatters, Ltd.) using "medium" assembly parameters for the non-classical cholera loci (maximum gaps per read = 15%, maximum gap size = 50, word length = 14, index word length = 12, ignore words repeated more than 10x, maximum mismatches per read = 30%, maximum ambiguity = 4), and "low" assembly parameters for the human loci (maximum gaps per read = 10%, maximum gap size = 3, word length = 24, index word length = 14, ignore words repeated more than 1x, maximum mismatches per read = 10%, maximum ambiguity = 4). Assemblies were collapsed for PCR duplicate reads using the *samtools* rmdup (-s) program.[11]

iii. *Ancient DNA Damage Assessment*. We subjected the post-trimmed human mitochondrial genome reads to analysis using the MapDamage (v.0.3.6) tool[12] (**Figure S4c**). Patterns of damage (C→T or G→A changes at the 5' and 3' fragment ends) typical to ancient DNA were seen, despite the differences in tissue, age, preservation method, etc. between this sample and other "typical" ancient DNA specimens.

C. Other Analyses

i. *Assessment of GIs Present in Strain RC27.* The downloaded contigs of the scaffold genome of strain RC27 (accession NZ_ADAI00000000.1) were aligned to the O395 chromosomes (NC_009456, NC_009457) using Geneious (5.6.5) highest sensitivity assembly parameters and allowing for mapping multiple best matches. **Table S7** shows the results for the PG-2 GIs of interest. Note that we did NOT investigate for the presence of any non-O395 genomic islands of interest, so it is possible/likely that RC27 does harbor additional GIs found in other strains.

III. SUPPLEMENTARY RESULTS

A. PA1849 vs O395

i. *CTX Prophages.* Reads support a possible tandem configuration of CTX elements on one OR both chromosomes (**Figure S8e**), as well as the immediate (**Figure S8a-d**) and more distant (not shown) chromosomal flanking regions, which demonstrate that the locations of the CTX element(s) are the same in PA1849 as in modern O395. In **Figure S8e**, only those reads not clearly supporting the other arrangements (as in **S8a-d**) are shown assembled to the CTX-CTX tandem span region from the B33 strain (accession GQ485644).[17] However, due to short fragment lengths in the historical specimen, we are unable to determine the exact configuration of CTX elements (**Figure S9**): a double tandem CTX prophage may exist on just one or both chromosomes (or, if not actually in the genome, in plasmid form).

Additionally, we have no evidence for the "truncated" CTX prophage that is seen adjacent to the full prophage on the large chromosome in modern classical strains. Instead, reads that map to this region only map to the immediate 5'/3' sequences that would correspond to a single full prophage (**Figure S10**).

ii. *"Superintegron".* Due to the repetitive nature of the integron cassettes, it is not possible to completely reconstruct such a locus with an ancient sample, due to the highly fragmented nature of the DNA molecules (as short fragment lengths precludes specific assignment of reads falling into repeat regions). While it is therefore not possible to reliably reconstruct the PA1849 integron gene order, we have here reported on the genic content of the integron vs. O395. **Table S8** highlights the PA1849 coverage, O395 %GC, and percentage of the O395 reference sequence covered by at least one read for many superintegron genes of interest: those gene cluster (i.e., groups of homologous genes) that form the "core" integron genome and those genes unique to O395, as outlined in reference 18. Overall the gene content of the integron appears very similar to O395, although as noted previously, PA1849 is missing GI-14 (a large section of the integron). All the "core" genes are apparently present, as are most of the O395 genes. Several (non-GI-14) O395-specific genes MAY be missing due to lower average coverage and percent of the reference covered than the genomic average; however, %GC may be contributing to these lower statistics, as most of these are slightly lower than average %GC (these genes of unknown presence are marked "Y?" in the **Table S8** "confirmed?" column).

iii. *CRISPR Region (GI-24).* Overall, the CRISPR genic region (falling in GI-24) appears to be largely similar to O395, with differences in the 'spacer' region. The CAS/CSE genic content is the same and appears the same order as O395 (*cas3', cse1, cse2, cas6e, cas7, cas5, cas1,* and *cas2*, followed by the repeat/spacer region).[19,20] The repeat/spacer region has a visibly punctuated coverage pattern that differs from the rest of the island due to the lack of coverage at the majority of the CRISPR spacers, indicating that PA1849 likely had different spacer content than O395. Unfortunately, it is not possible to reconstruct the content with these types of data, due to short fragment lengths and lack of proper enrichment targets. **Figure S11** visually demonstrates the coverage pattern across the CRISPR repeat/spacer region.

iv. *Other Loci.* Comparing other genes of interest between PA1849 and modern classical indicate that the strains are extremely similar.[21] The PA1849 strain possesses the classical (deletion) variant of *hlyA,* hemolysin at all covered positions. The hemagglutinin protease (Hap) sequence is identical to the O395 reference at all covered positions. The mannose-sensitive hemagglutinin (MSHA) genes (such as *mshA*) are present and identical to O395 at all covered positions, except for 1 SNP in *mshN* (2,973,101). Regarding the antibiotic resistance element SXT, we can report that (as expected) there is no evidence of its presence because the reads falling across the SXT integration site at the 5' end of *prfC* (peptide chain release factor 3)[22] are identical to the O395 reference sequence, which does not contain SXT. SXT is currently only found among recent El Tor and related strains, and was not included as one of the several "non-classical" loci on this enrichment array (which was not intended to be an exhaustive list of all possible 'pan-genomic' cholera content).

B. PA1849 vs. Other Strains (Regions Not Expected to Occur in the Classical Strain)

All regions returned the expected outcomes, which were: low coverage and percent of the reference covered for those regions absent from the classical genome; the correct (classical) variant for those regions with significant divergence to classical (see **Table S4**).

C. Human Loci

i. *Mitochondrial Genome*. The assembly relative to the revised Cambridge Reference Sequence yielded 37 SNPs relative to the reference, with high unique coverage and variant frequencies. Mitotool.org returns the haplogroup as L3d1b3, which is consistent with the L3d1b3 SNPs recorded at phylotree.org (both accessed 2012). Coverage information can be found in **Table S5**; SNP calls can be found in **Table S6**.

ii. *Amelogenin Genes.* The coverage of reads across the human AMELX and AMELY genes was highly variable (see **Table S5**). Low coverage regions tend to be low in GC% (see **Figure S12**). Due to the extreme coverage disparities and the likely inclusion of many non-amelogenin reads in this assembly (at conserved regions), it is not possible to directly reconstruct the sequence of these genes. However, several reads BLAST exclusively to the human AMELY gene (BLASTN; default parameters), suggesting that the individual was indeed a male, though more work would be necessary to be conclusive.

IV. SUPPLEMENTARY FIGURES

Figure S1. qPCR amplification curves for 58bp *ompW* assay

One replicate of each standard of known copy number (10K, 1K, 100, and 10) are shown in black. PCR blanks are shown in grey. Library amplifications (1 in 100 dilution; 0.01x) are shown in green (3090.13 in dark green, all other blanks in light green). Indexing amplifications (0.01x) are shown in blue (3090.13 in light blue, all other blanks in very light blue). Re-amplification blanks (0.01x) are shown in violet. The threshold RFU (relative fluorescence units) level to calculate Cq threshold was automatically determined at 69.01 (BioRad CFX Manager 3.0 software). Cq values shown below for the positive template amplifications were as follows: 3090.13 library (0.01x) replicates (35.35, 36.29), 3090.13 indexed library (0.01x) replicates (35.58, 35.75). All other templates tested (blanks) were negative for the 58bp *ompW* product after 50 cycles.

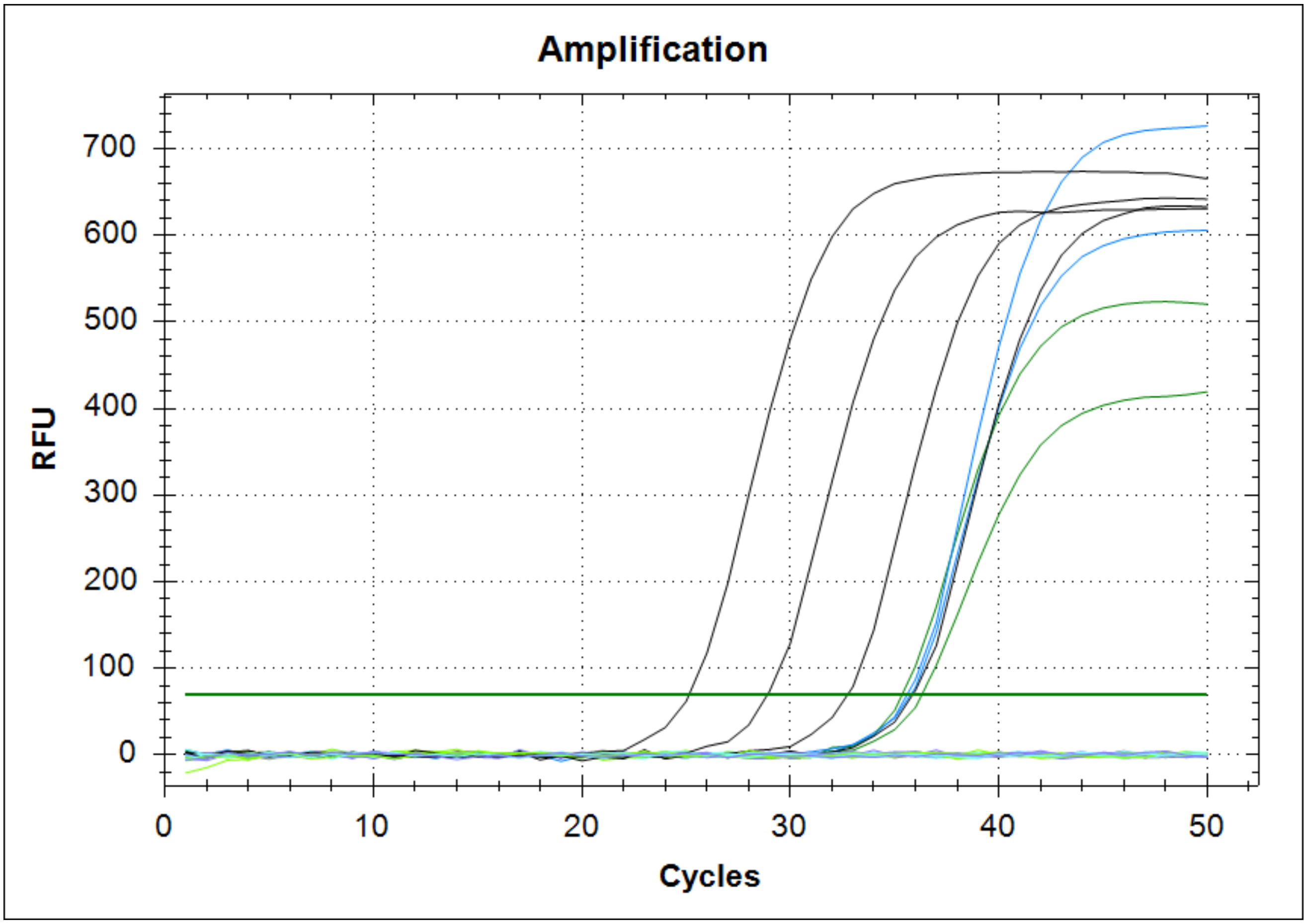


Figure S2. Fragment length distributions (FLDs)

The *V. cholerae* chromosomal FLDs reflect all unique reads. The human mitochondrial FLD reflects the 14,731 collapsed unique reads (samtools rmdup) from a 33,000 random raw read subset (out of 19.1 million total raw reads).

Figure S2a. FLD: V. cholerae large chromosome

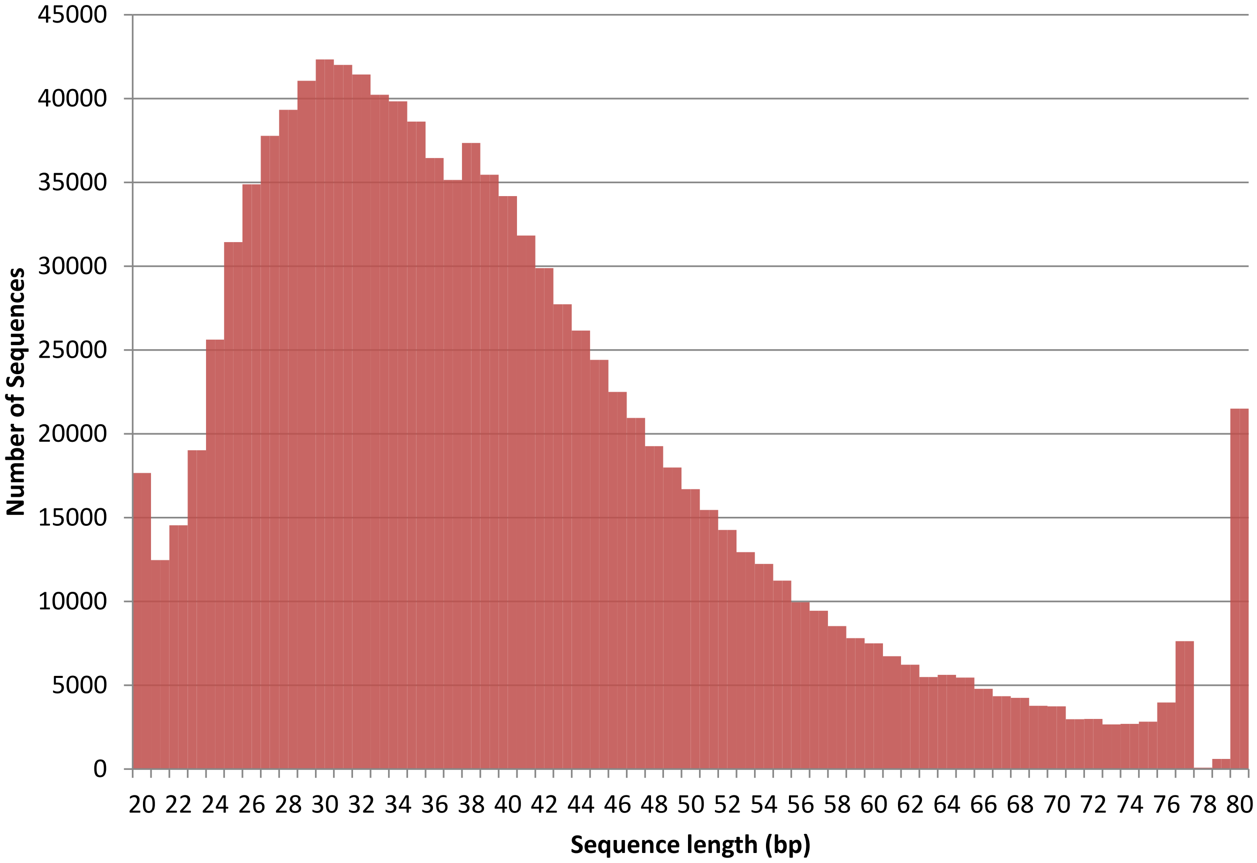


Figure S2b. FLD: V. cholerae small chromosome

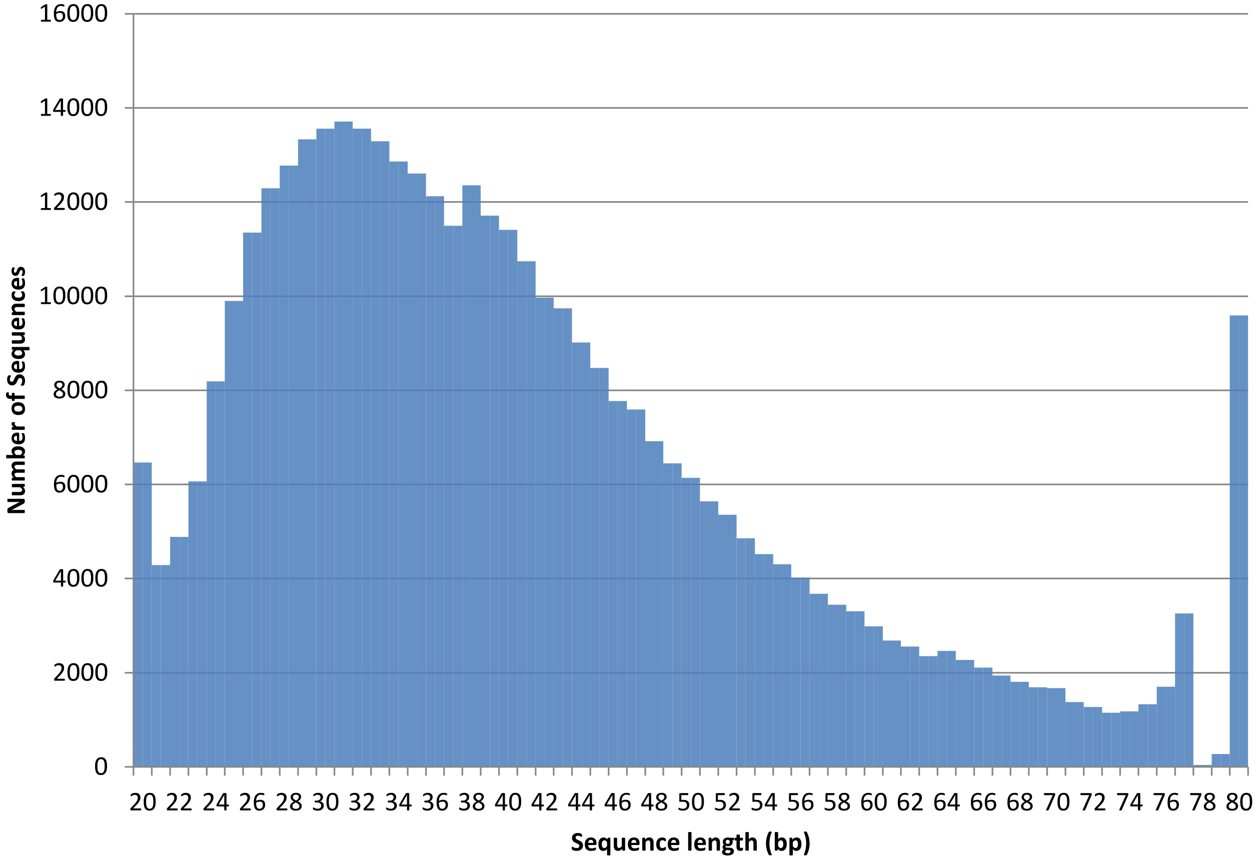


Figure S2c. FLD: Human mtDNA

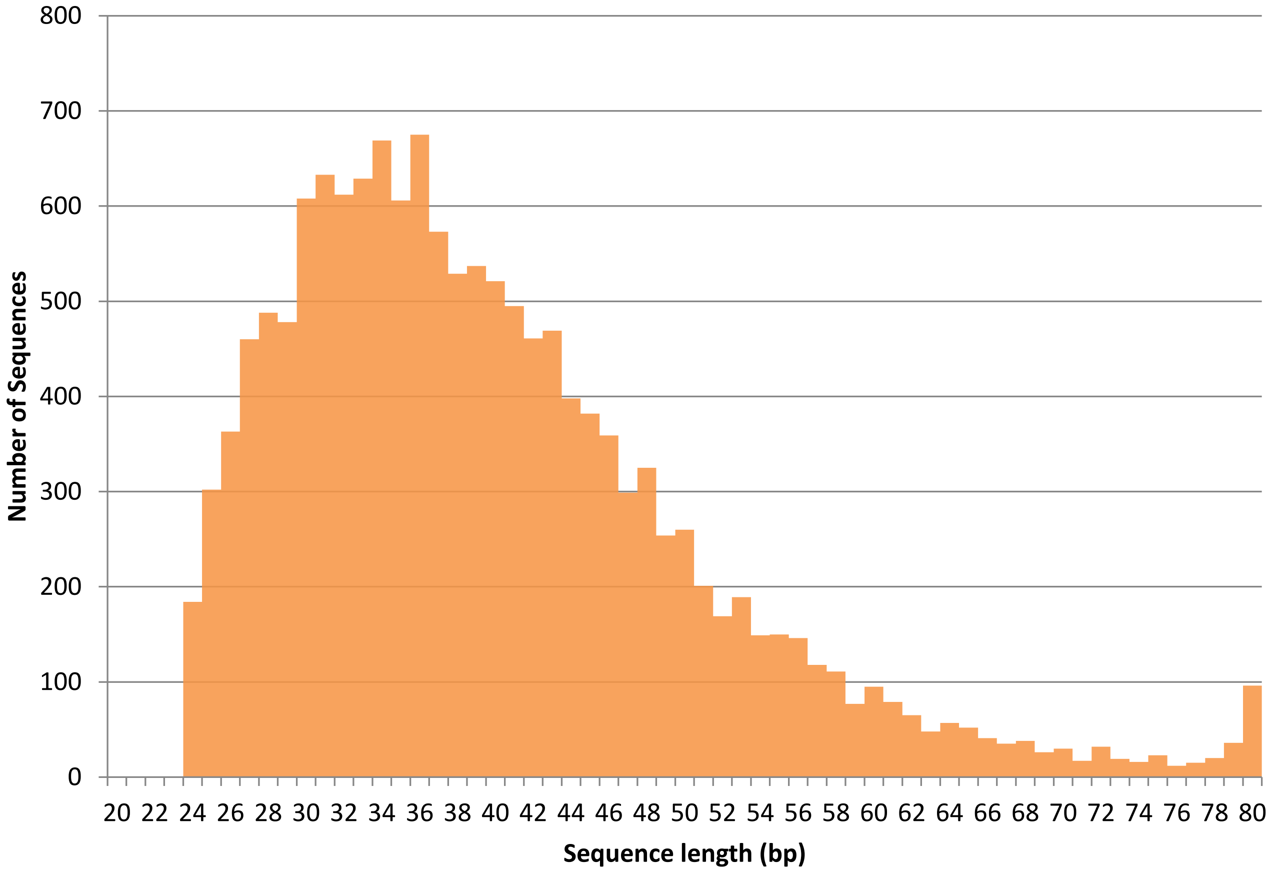


Figure S3. PA1849 coverage vs. GC% (100bp window)

Plot of average unique PA1849 coverage vs GC% of the O395 reference genome across 10bp consecutive windows. To reduce complexity, only every 10th data point is considered here. Coverage axis is scaled logarithmically. The logarithmic trendline and equation of best fit is shown (R^2 = 0.25).

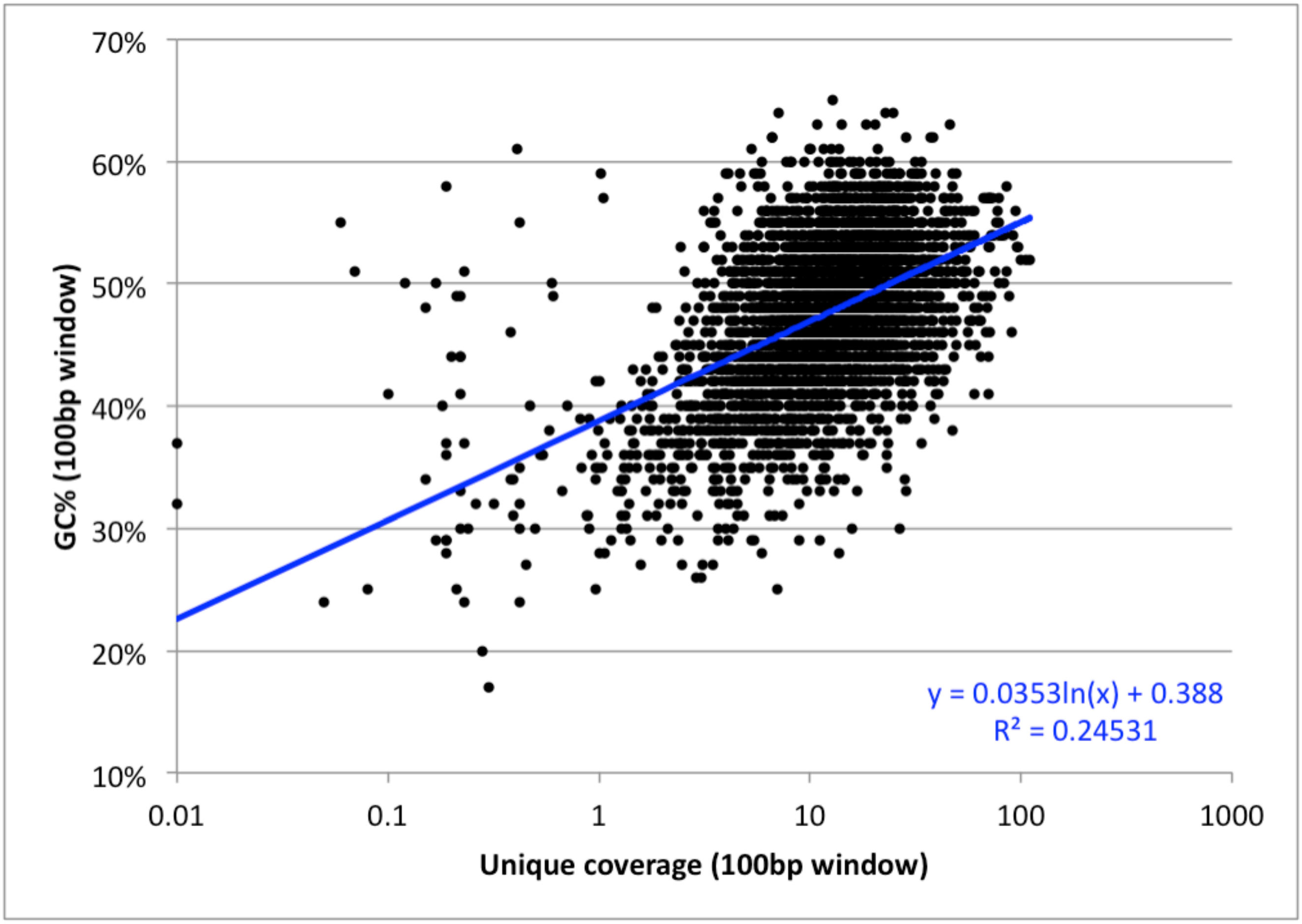


Figure S4. mapDamage graphs

The *V. cholerae* chromosomal mapDamage profiles reflect all unique reads. As in Figure S2, the human mitochondrial mapDamage profile reflects the 14,731 collapsed unique reads (samtools rmdup) from a 33,000 random raw read subset (out of 19.1 million total raw reads).

Figure 4a. mapDamage: V. cholerae large chromosome

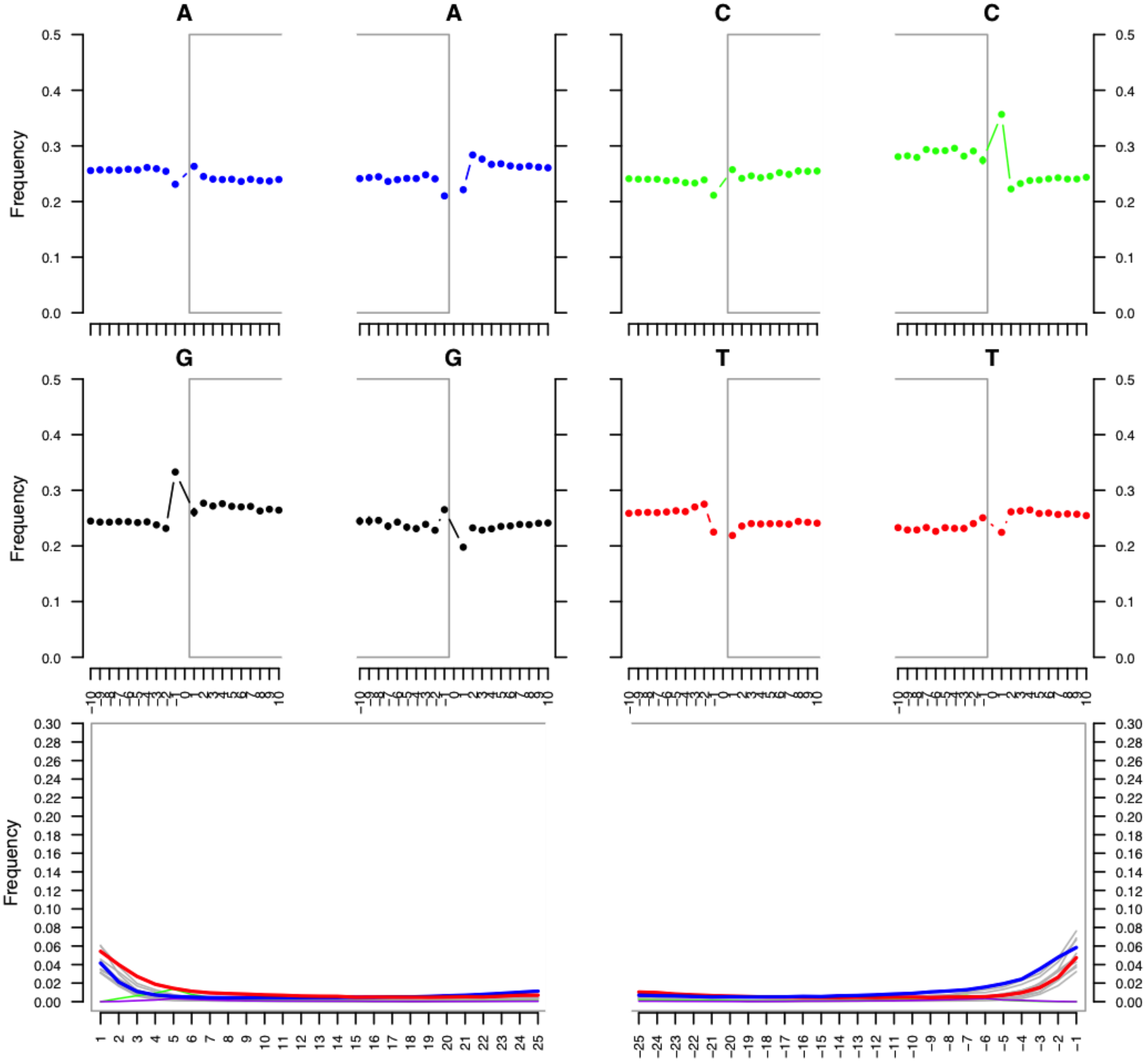


Figure S4b. mapDamage: V. cholerae small chromosome

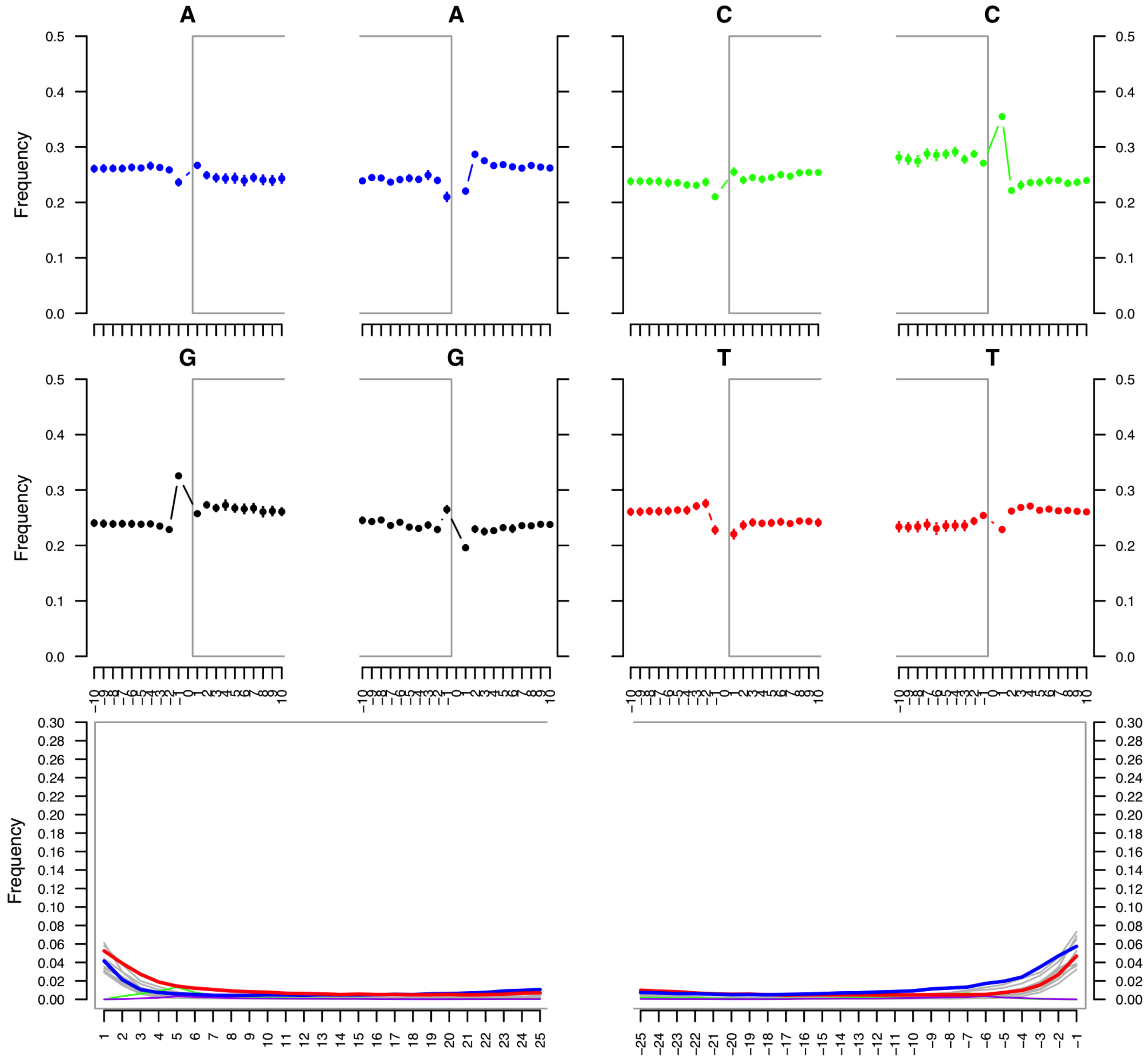


Figure S4c. mapDamage: Human mtDNA.

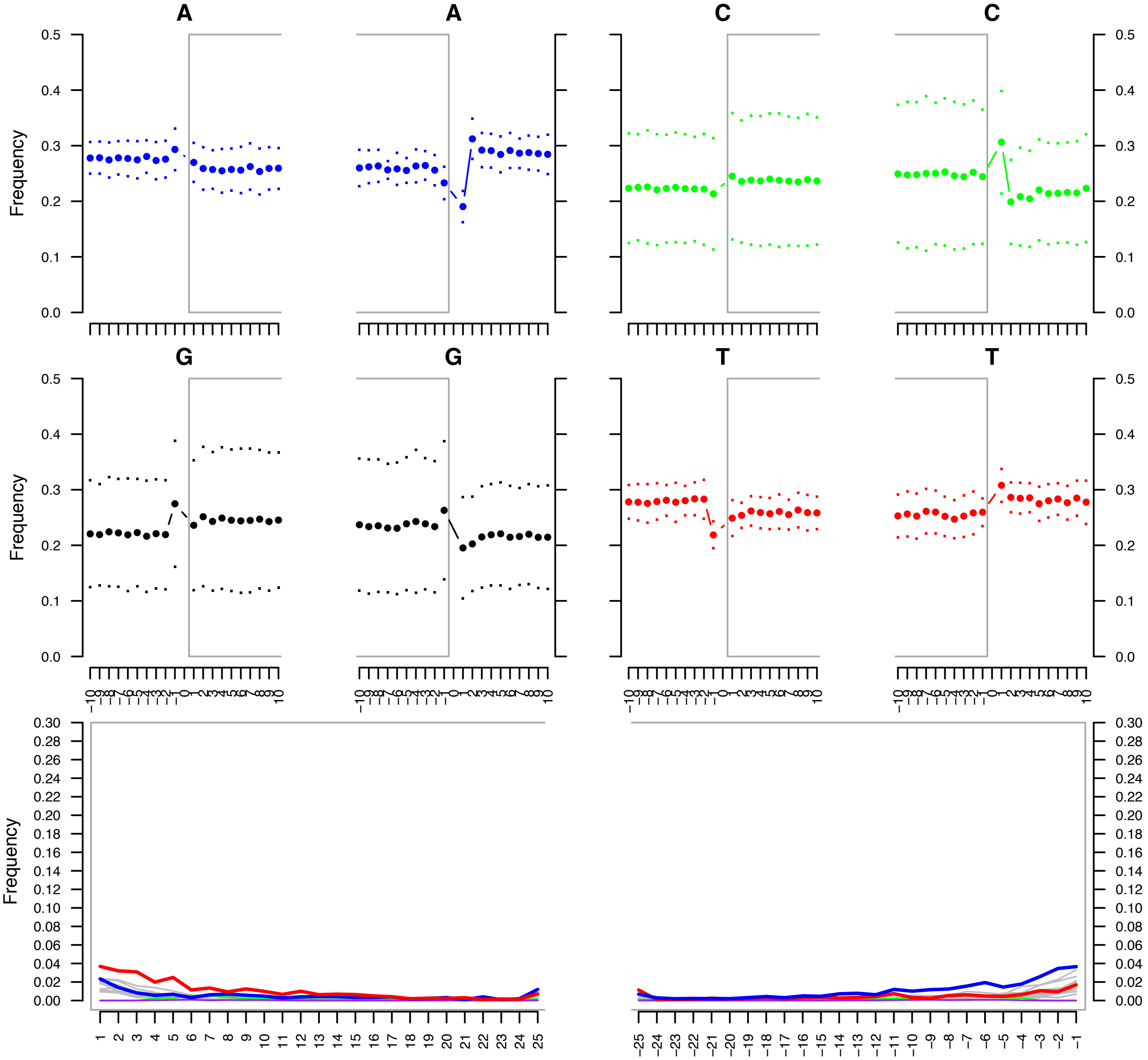


Figure S5. Map of known geographic strain locations

Map of all known geographic locations of strains used in the evolutionary analysis, grouped by geographic region. (Note that some of the strains used for the phylogenetic analysis were of unknown provenience.) Size of the nodes corresponds to number of samples from that region (from west to east: Pennsylvania USA n=1, southern North America n =2, Brazil n=1, Sweden n=1, Red Sea region n =4, Bay of Bengal region n = 14, Southeast Asia n = 5). Colors correspond to the clade colors used in the phylogenetic tree in the main text, **Figure 3**.

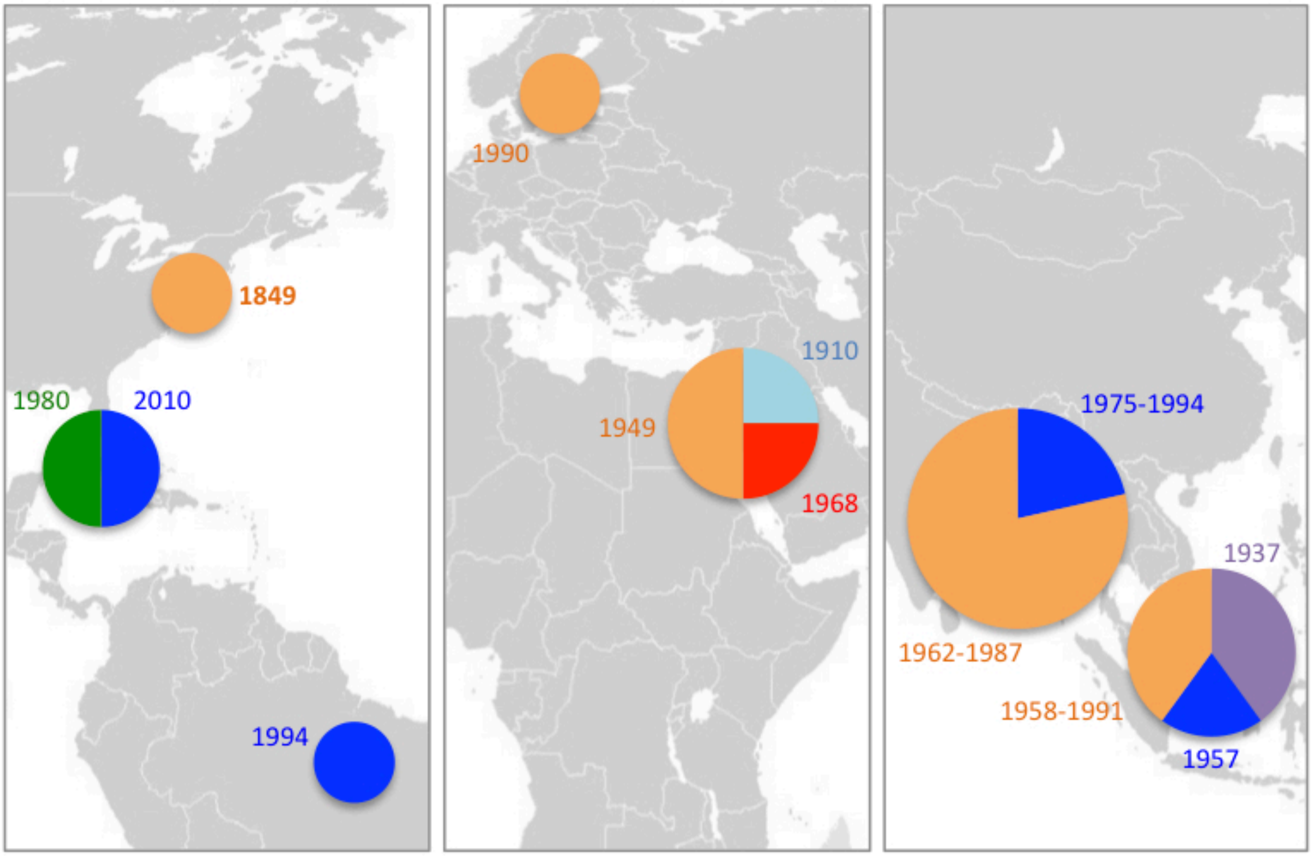


Figure S6. Maximum likelihood (ML) phylogeny of 32 strains of *V. cholerae* excluding recombinant regions. This tree is the same as that shown in **Figure 3** in the main text, with the exception that recombinant regions (detected by RDP4; see above) have been excluded. The tree is mid-point rooted for clarity and all horizontal branch lengths are drawn to a scale of nucleotide substitutions per SNP site.

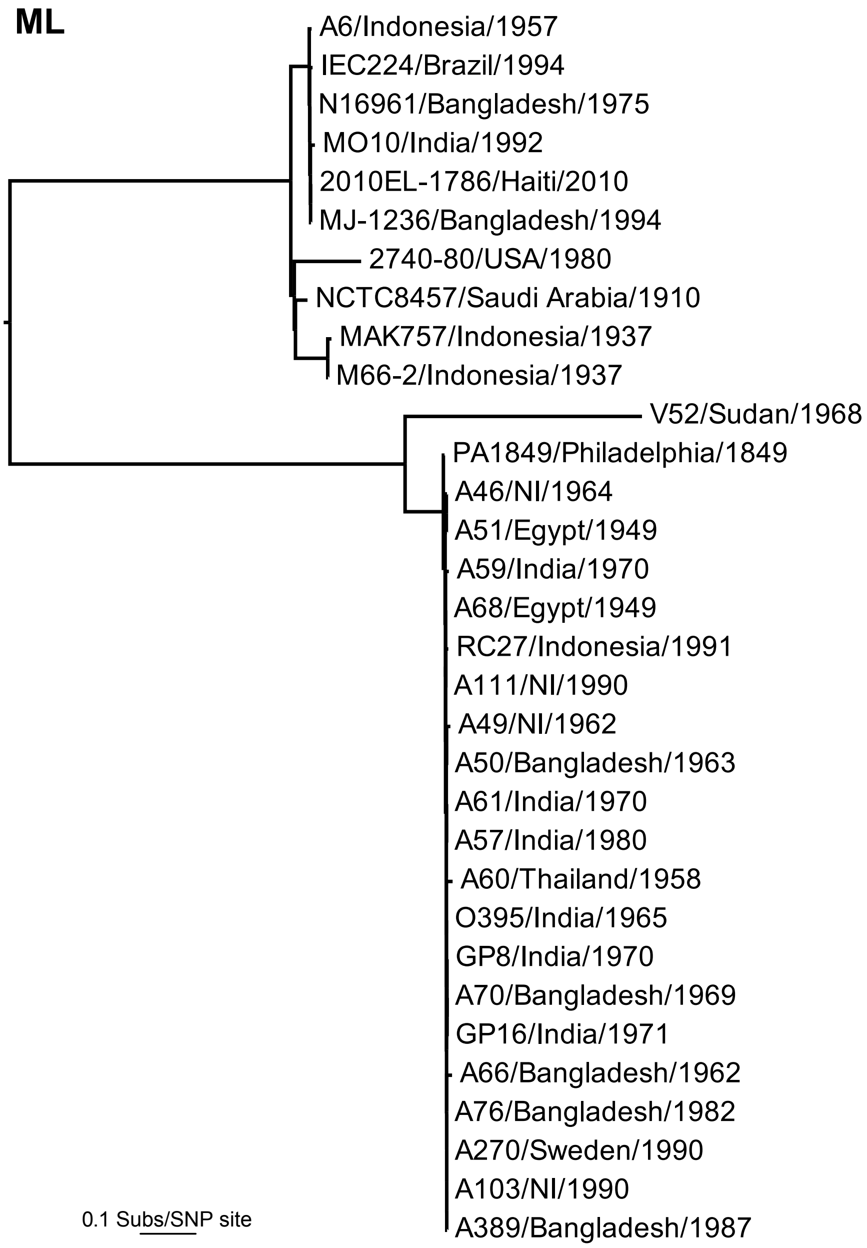


Figure S7. Maximum parsimony phylogeny of 32 strains of *V. cholerae* excluding recombinant regions.

To improve visual resolution of the branching order, branch lengths are not drawn to scale. The tree is mid-point rooted for clarity only.

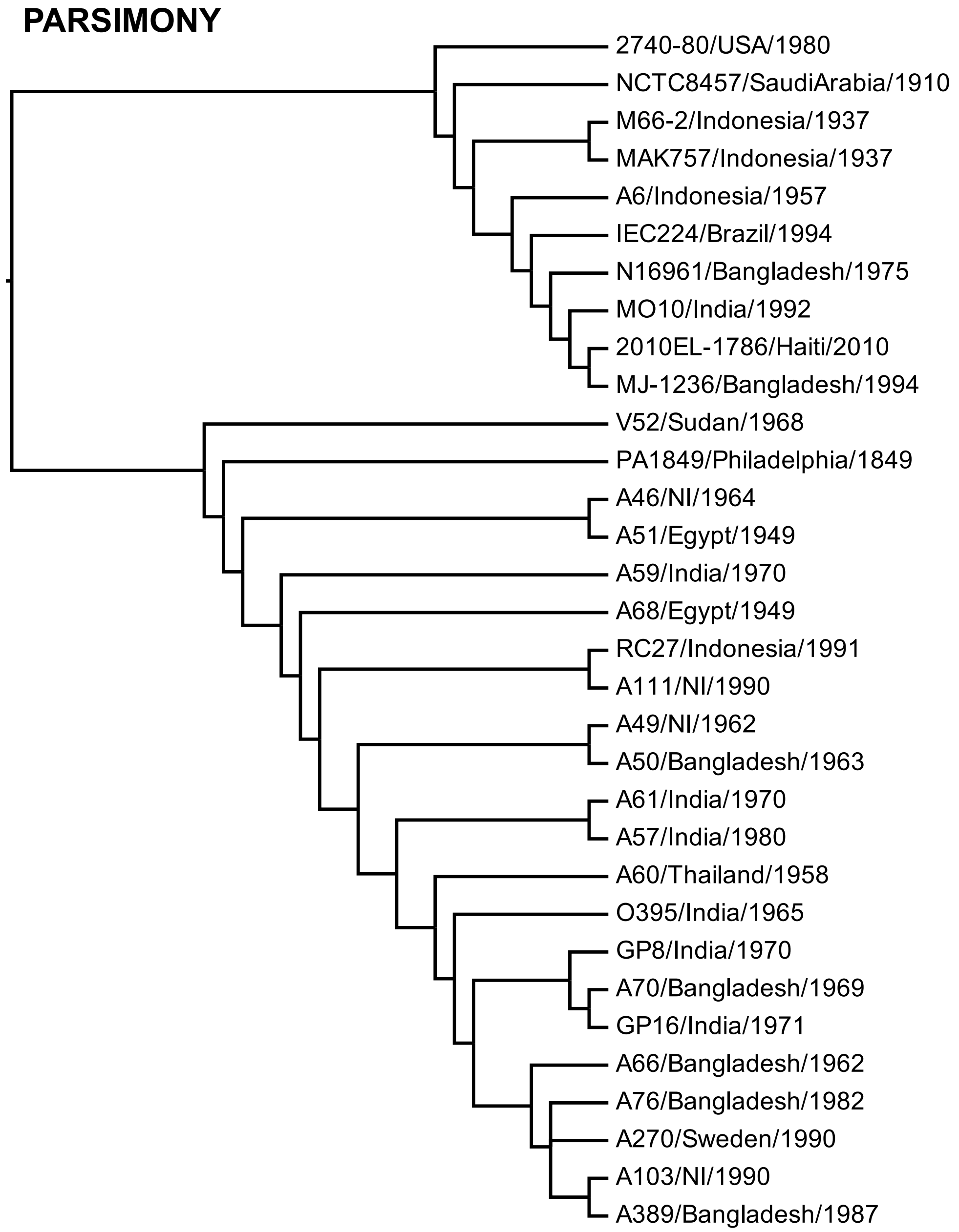
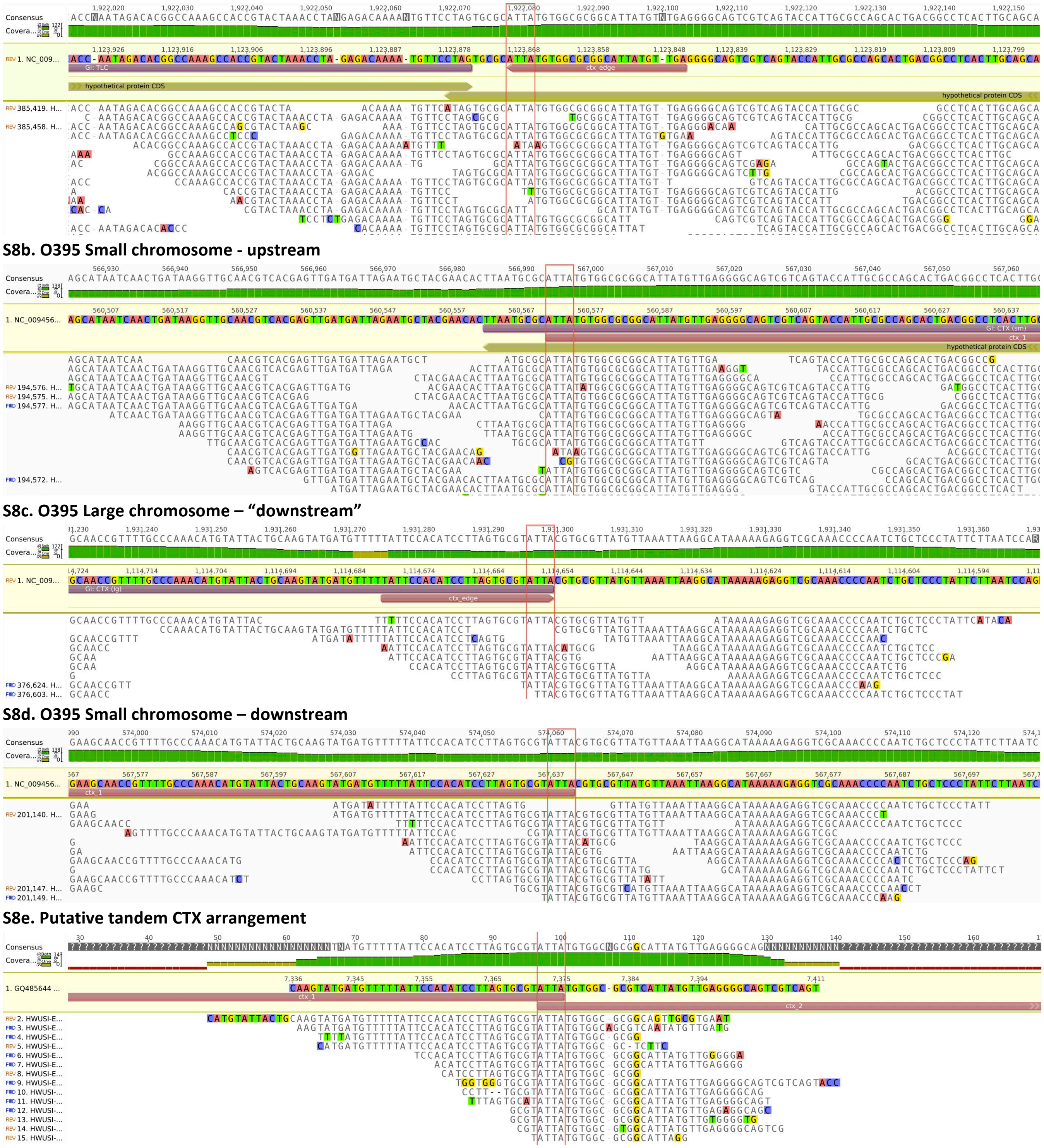


Figure S8. Assemblies supporting CTX prophage arrangements

Assemblies showing the “upstream” and “downstream” edges of CTX prophage on both chromosomes (vs. strain O395; **a-d**) and possible PA1849 tandem CTX arrangement (vs. strain B33, accession GQ485644; **e**). CTX orientation is shown as in O395 small chromosome. Both chromosomes share the same immediate “downstream” flanking at this scale (**c & d**). Red box highlights the ‘ATTA’ motif at the CTX edges in each assembly. Assembly, graphs, and annotations visualized using Geneious (v.6.1.6, Biomatters, Ltd).



S8a. O395 Large chromosome – “upstream”

S8b. O395 Small chromosome - upstream

S8c. O395 Large chromosome – “downstream”

S8d. O395 Small chromosome – downstream

S8e. Putative tandem CTX arrangement

Figure S9. Possible PA1849 CTX region structural configurations

O395 CTX region arrangement has full and truncated CTX prophage sequences on the large chromosome adjacent to the TLC element, and a full CTX prophage on the small chromosome. Based on read data and assemblies (see text and Figure S7), PA1849 has at least a CTX prophage tandem arrangement on one or both chromosomes, and no evidence of a truncated/full CTX prophage structural arrangement. One of three possible chromosomal CTX arrangements exists: a single CTX on the large and tandem on the small, tandem on the large and single on the small, or tandem on both chromosomes. (It is also possible (not shown) that the tandem orientation exists only in plasmid format.) Because of the short fragment lengths in the historical specimen, we are unable to exactly determine the original configuration.

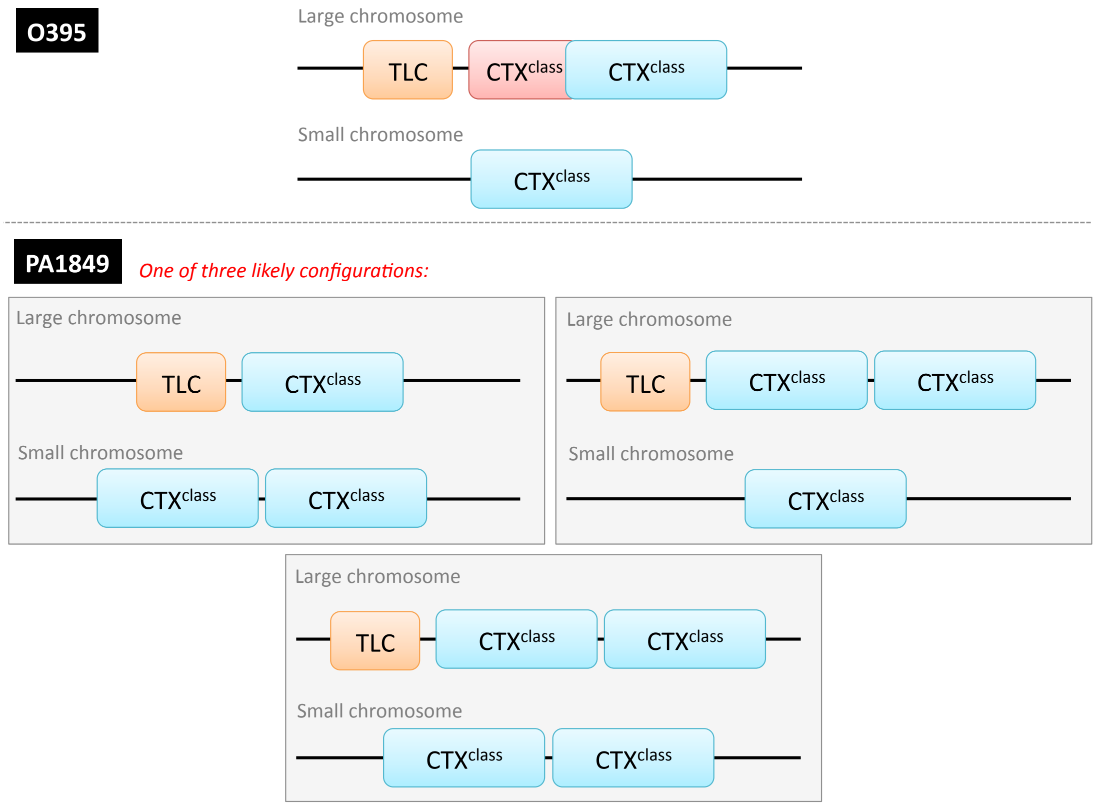


Figure S10. No evidence for truncated CTX prophage

All unique PA1849 reads mapping to the ‘duplicated’ rst region (truncated CTX prophage adjacent to full CTX prophage) on the large chromosome of O395 are shown. A red box outlines the overlap ‘A’ between rst regions. No reads span this intersection, indicating that PA1849 lacks the truncated CTX prophage. Assembly, graphs, and annotations visualized using Geneious (v.6.1.6, Biomatters, Ltd).

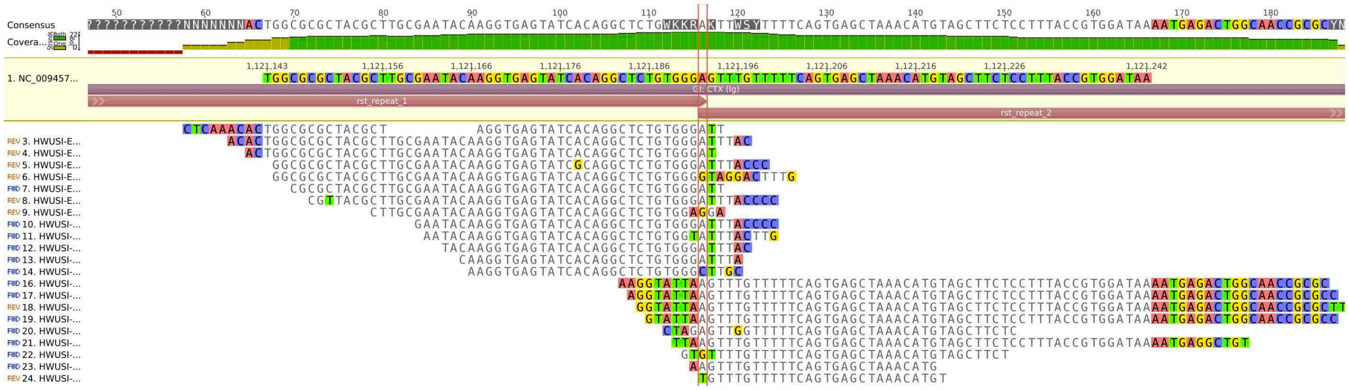


Figure S11. PA1849 coverage across the O395 CRISPR region of GI-24.

Graphical representation of PA1849 reads assembled to the O395 GI-24 CRISPR region, highlighting (left to right) genes *cas3’, cse1, cse2, cas6e, cas7, cas5, cas1, cas2*, and repeat-spacer region (gene terminology from reference 20). Coverage of the repeat-spacer region is high in the repeats but almost zero in the spacers, indicating that PA1849 does not contain similar spacer content to O395 (with a few exceptions). Assembly, graphs, and annotations visualized using Geneious (v.6.1.6, Biomatters, Ltd).

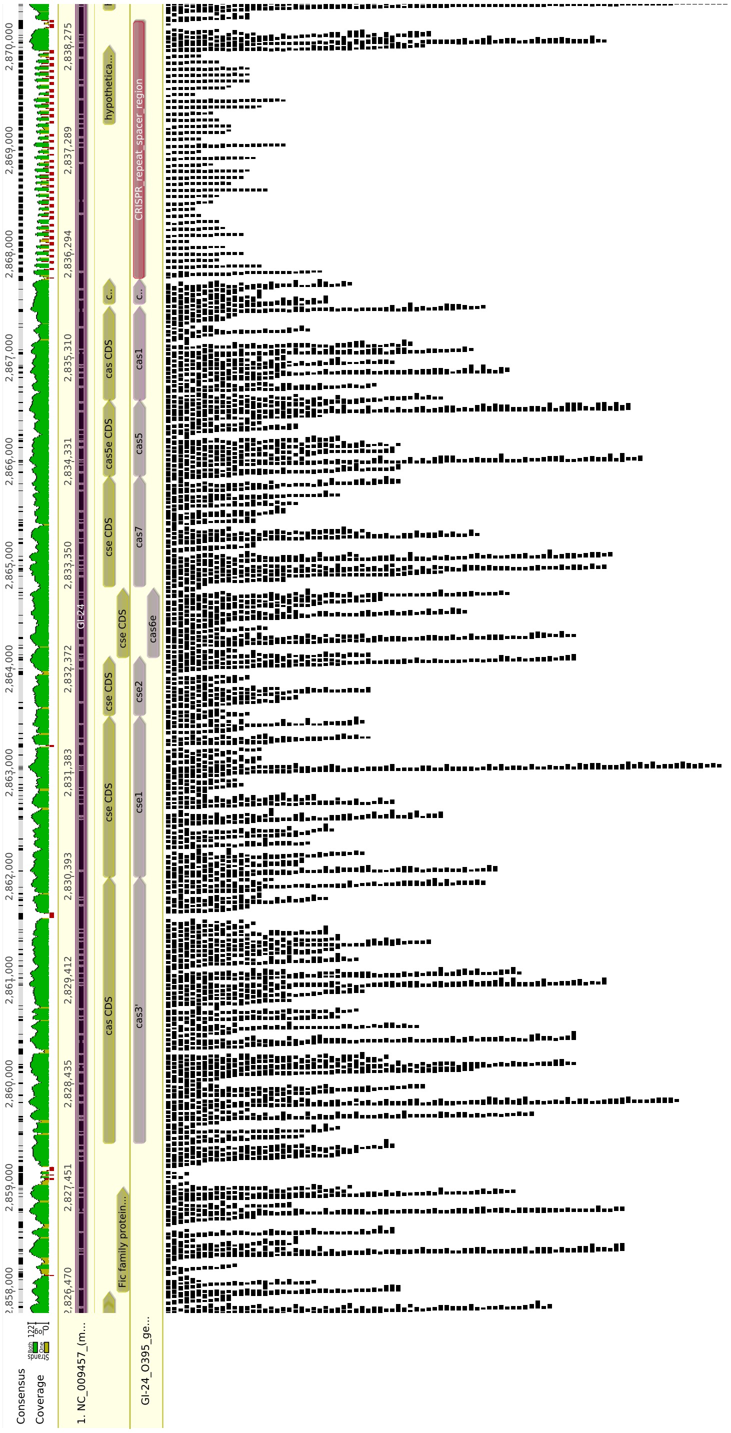
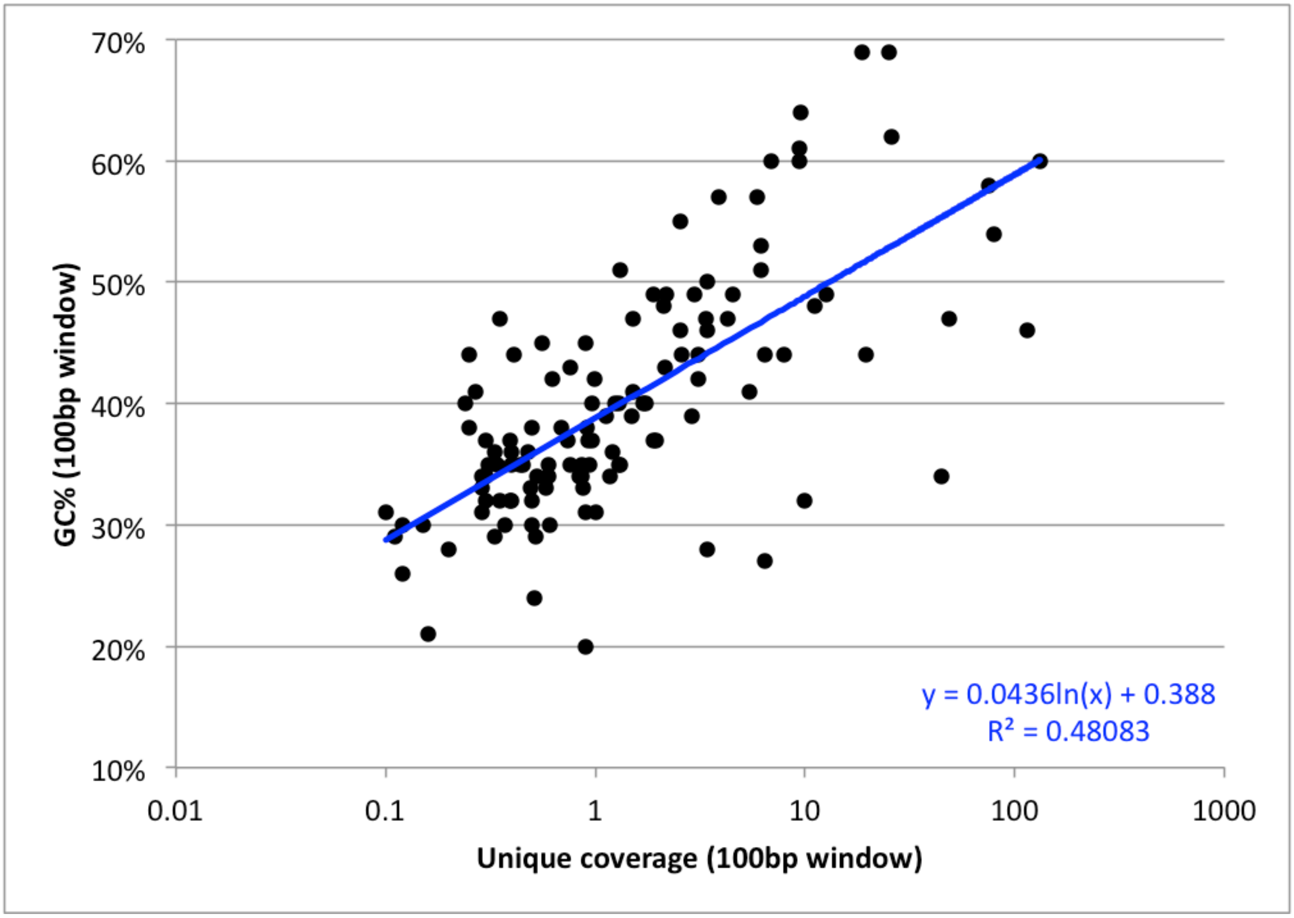


Figure S12. Human amelogenin X and Y coverage vs. GC% (100bp window)

Plot of average unique 3090.13 coverage across the human AMELX and AMELY genes vs GC% of the reference sequence across 100bp consecutive windows. Coverage axis is scaled logarithmically. The logarithmic trendline and equation of best fit is shown (R^2 = 0.48).



V. SUPPLEMENTARY TABLES

Table S1. Sequences included on enrichment array

#	Accession	Description	Length	Position	Tiling	# probes	Ref
		CHOLERA - EXPECTED					
1	NC_009457	"Vibrio cholerae O395 chromosome 2, complete sequence" (large)	3,024,069	-	5	603,469	
		recA	1265	65,378-66,642	5 (2 offset)	242	
		RTX	35735	1,103,698-1,139,432	5 (2 offset)	7,136	23
		VPI	42290	363,273-405,562	5 (2 offset)	8,447	24
		VPI2	58540	1,448,788-1,507,327	5 (2 offset)	11,697	25
2	NC_009456	"Vibrio cholerae O395 chromosome 1, complete sequence" (small)	1,108,250	-	5	220,898	
		CTX	7540	560,458-567,997	5 (2 offset)	1,497	26
		hlyA	1705	986,188-987,892	5 (2 offset)	330	
		Integron (partial)	5830	911,858-917,687	5 (2 offset)	1,155	
		ompW	855	404,753-405,607	5 (2 offset)	160	
		CHOLERA - NOT EXPECTED or DIFFERENT VARIANTS					
	NC_002505	Vibrio cholerae O1 biovar El Tor str. N16961 chromosome I, complete genome	-	-	-	-	
3		VSP-I	16,946	173,969-190,914	2	8,443	27
4		RS1 prophage	3,182	1,563,785-1,566,966	2	1,561	28
5		recA (El Tor variant)	1,239	574,522-575,760	2	591	
6		VC0514-VC0516	5,048	545,074-550,121	2	2,494	27
7		tcpA (El Tor variant)	675	890,449-891,123	2	309	
8		VSP-II	27,790	522,397-550,186	2	13,865	27
9		rtxC (VC1449-VC1450)	1,190	1,548,919-1,550,108	2	565	27
	NC_002506	N16961 El Tor chr 2 (sm)	-	-	-	-	
10		hlyA (El Tor variant)	60	238,297-238,356	2	1	
11		VCA0300	830	315,211-316,040	2	385	27
12	AB012956	O-antigen synthesis, strain MO45 [O139]	35,806	-	2	17,873	29
		HUMAN – EXPECTED and/or UNKNOWN					
13	RefSeq accession NG_012040	Homo sapiens amelogenin, X-linked (AMELX), RefSeqGene on chromosome X	8,059	4,681-12,739	4	2001	
14	RefSeq accession NG_008011	Homo sapiens amelogenin, Y-linked (AMELY), RefSeqGene on chromosome Y	8,759	4,737-13,495	4	2176	
15	NC_012920	Homo sapiens mitochondrion, complete genome	16,569	-	4	4129	30
		OTHER - NOT DISCUSSED					
16	-	Additional non-project probeset	48,517	-	-	13,729	

Table S2. Genomes used for phylogenetic comparison

Strain	Description	Date	Location	Accession (lg, sm chr)	Reference
NCTC 8457	O1 Inaba - El Tor, nonpandemic	1910	Saudi Arabia	NZ_AAWD00000000.1	NCBI
M66-2	O1, pre-7th pandemic	1937	Indonesia	NC_012578.1, NC_012580.1	31
MAK 757	O1 Ogawa - El Tor	1937	Indonesia	NZ_AAUS00000000.2	NCBI
A51	O1 Ogawa - classical (Cairo 50)	1949	Egypt	ERS013165 [SRA dataset]	32
A68	O1 Inaba - classical (Cairo 48)	1949	Egypt	ERS013171 [SRA dataset]	32
A6	O1 – El Tor	1957	Indonesia	ERS013246 [SRA dataset]	32
A60	O1 Inaba - classical	1958	Thailand	ERS013168 [SRA dataset]	32
A49	O1 Inaba - classical	1962	unknown	ERS013161 [SRA dataset]	32
A66	O1 Inaba - classical	1962	Bangladesh	ERS013170 [SRA dataset]	32
A50	O1 Ogawa - classical	1963	Bangladesh	ERS013164 [SRA dataset]	32
A46	O1 Ogawa - classical	1964	unknown	ERS013160 [SRA dataset]	32
O395	O1 Ogawa - classical	1965	India	NC_009457.1, NC_009456.1	NCBI
V52	O37, clinical	1968	Sudan	NZ_AAKJ00000000.2	NCBI
A70	O1 Inaba - classical (G28190)	1969	Bangladesh	ERS013162 [SRA dataset]	32
A59	O1 Inaba - classical	1970	India	ERS013167 [SRA dataset]	32
A61	O1 Inaba - classical	1970	India	ERS013169 [SRA dataset]	32
GP8	O1 Inaba - classical	1970	India	ERS013128 [SRA dataset]	32
GP16	O1 Inaba - classical	1971	India	ERS013136 [SRA dataset]	32
N16961	O1 Inaba - El Tor	1975	Bangladesh	NC_002505.1, NC_002506.1	33
2740-80	O1 Inaba - El Tor, environmental	1980	USA	NZ_AAUT00000000.1	NCBI
A57	O1 Ogawa - classical (U10198)	1980	India	ERS013166 [SRA dataset]	32
A76	O1 Inaba - classical (X19850)	1982	Bangladesh	ERS013163 [SRA dataset]	32
A389	O1 Inaba - classical (VM11647)	1987	Bangladesh	ERS013203 [SRA dataset]	32
A111	O1 Inaba - classical (V591)	1990	unknown	ERS013176 [SRA dataset]	32
A279	O1 Inaba - classical (K216/92)	1990	Sweden	ERS013197 [SRA dataset]	32
A103	O1 Inaba - classical (V584)	1990	unknown	ERS013172 [SRA dataset]	32
RC27	O1 - classical	1991	Indonesia	NZ_ADAI00000000.1	NCBI
MO10	O139	1992	India	NZ_AAKF00000000.3	NCBI
MJ-1236	O1 Inaba - El Tor (Matlab variant)	1994	Bangladesh	NC_012668.1, NC_012667.1	34
IEC224	O1 - El Tor	1994	Brazil	NC_016944.1, NC_016945.1	35
2010EL-1786	O1 - El Tor	2010	Haiti	CP003069.1, CP003070.1	36

Table S3. Results of genomic islands of interest

Class	#§	Chr	Description§	O395 position	%GC	Avg unique coverage	% reference covered @1X	Confirmed?
“PG”§	O1	lg	O1-antigen region	2,771,546-2,795,569 #	40.4	10.0	87.10%	Y
	GI-1	lg	Motility and chemotaxis	1,043,291-1,055,071	43.4	14.4	97.80%	Y
	GI-2	lg	Oxidative stress response	1,248,952-1,252,534	44.1	12.9	93.00%	Y
	GI-3	lg	Membrane proteins	1,435,685-1,440,550	44.3	17.3	98.80%	Y
	GI-4	lg	Carbohydrates (PTS system)	1,514,201-1,523,686	46.4	18	98.00%	Y
	GI-5	sm	Site-specific DNA-methyltransferase	1,005,832-1,011,520	36	11.8	76.00%	Y*
	GI-6	sm	Putative prophage	919,877-925,639	41.6	16.7	94.20%	Y
	GI-7	sm	Sodium-solute symporter/Sugar transporter	664,079-669,638	42.9	14.2	96.20%	Y
	GI-8	sm	Transposable element	486,113-489,350	41.3	26.1	97.40%	Y
	GI-9	sm	Autolysin sensor kinase/ABC-type transport system	417,216-423,212	48.2	19	99.70%	Y
	GI-10	sm	Integrase/Non-hemolytic enterotoxin lytic component L1	388,323-393,223	40.3	11.9	91.00%	Y
	VPI-1	lg	Vibrio pathogenicity island-1 (VPI-1)	355,973-404,865	35.9	10.1	81.40%	Y*
	VPI-2	lg	Vibrio pathogenicity island-2 (VPI-2)	1,449,205-1,505,734	41.3	21	97.30%	Y
Post “PG-2”§	GI-11	sm	Kappa prophage	1012862-1045566	48	0.6	26.30%	**N**
	GI-14	sm	Hypothetical proteins	814223-832392	37.7	0.6	18.00%	**N**
	GI-21	lg	Mu-like prophage	689569-722653	50.4	0.4	19.40%	**N**
		lg	Mu-like prophage	766615-799585	50.5	0.4	19.20%	**N**
	GI-23	lg	Putative prophage	2499645-2522558	39.8	15.6	97.90%	Y
	GI-24	lg	Putative prophage (CRISPR-associated proteins)	2825203-2840525	46.1	21.8	96.70%	Y
Other	CTX (lg)	lg	Cholera toxin prophage (CTX)	1,114,655-1,123,257	41.8	64.6**	99.70%	Y
	CTX (sm)	sm	Cholera toxin prophage (CTX)	560,562-567,519	42.7	67.0**	99.60%	Y
	TLC	lg	Cryptic plasmid linked to the CTX prophage	1,123,876-1,138,038	45	41.8**	99.80%	Y

§ - From reference 34; # - From reference 37

* - Low %GC likely contributes to the relatively lowered coverage and percent of reference covered at these loci.

** - Relatively higher coverage at these loci is possibly attributable to these loci being repeat regions of uncertain copy number in the ancient strain.

Table S4. Results of non-classical regions included on enrichment array

Accession	Description	Total length included on array	%GC	# reads assembled (raw)	# reads assembled (unique)	Avg unique coverage	% reference covered @1X	Hypothesized or expected presence or type	Confirmed?
NC_002505	*N16961 large chr*	-	-			-	-	-	-
	VSP-I	16,946							
		14,038 (VC0175-0185 only)	38.5	17,533	2,686	0.5	21.2%	Absent	Y
	VSP-II	27,790							
		26,866 (VC0490-0516 only)	39.9	4,633	1,016	0.8	23.9%	Absent	Y
	VC0514-VC0516	5,048	35.1	94	53	0.2	15.2%	Absent	Y
	rtxC (VC1449-VC1450)	1,190	42.1	125	45	0.9	28.8%	Absent	Y
	RS1 prophage (El Tor variant)	3,182	41.7	49,975	3,013	49.7	77.4%	Classical variant rstR; lacking rstC	Y
	tcpA (El Tor variant)	675	43.0	392	100	7.0	34.8%	Classical variant	Y
	recA (El Tor variant)	1,239	45.6	4,703	993	37.1	95.8%	Classical variant	Y
NC_002506	*N16961 sm chr*	-	-			-	-	-	-
	VCA0300	830							
		630 (VCA0300 only)	39.1	1,279,820	389	4.6	27.8%	Absent	Y
	hlyA (El Tor variant)	60	46.8	77	14	6.9	100%	Classical variant	Y
AB012956	***MO45 O-antigen synthesis genes (O139)***	35,806	40.4	7,236	1,591	1.7	25.6%	Absent	Y

Table S5. Results of human loci included on enrichment array

Accession	Description	Total length included on array	%GC	# reads assembled (raw)	# reads assembled (unique)	Avg unique coverage	% reference covered @1X	Relevant findings
NG_012040	Homo sapiens amelogenin, X-linked (AMELX), RefSeqGene on chromosome X	8,059	37.6	1,347	412	2.0	39.2%	X chr present
NG_009011	Homo sapiens amelogenin, Y-linked (AMELY), RefSeqGene on chromosome Y	8,759	36.6	1,403,301	1,235	14.1	41.5%	Y chr present (?)
NC_012920	Homo sapiens mitochondrion, complete genome	16,569	44.4	19,160,771	33,151	149.3	100%	Haplogroup L3d

Table S6. Human mitochondrial genome SNPs (relative to rCRS, NC_012920)

SNP #	rCRS position	rCRS base	3090.13 base	Unique coverage	% variant frequency
1	73	A	G	149	95.30%
2	146	T	C	140	94.30%
3	152	T	C	136	98.50%
4	263	A	G	128	99.20%
5	750	A	G	150	100.00%
6	921	T	C	151	98.00%
7	1438	A	G	138	97.80%
8	2706	A	G	138	95.70%
9	4769	A	G	132	98.50%
10	4937	T	C	161	95.70%
11	5046	G	A	128	100.00%
12	5147	G	A	150	96.00%
13	6680	T	C	138	97.80%
14	7028	C	T	143	100.00%
15	7424	A	G	152	93.40%
16	8618	T	C	155	94.20%
17	8701	A	G	144	99.30%
18	8860	A	G	159	95.00%
19	9540	T	C	155	95.50%
20	10398	A	G	134	97.00%
21	10694	A	T	142	97.90%
22	10873	T	C	157	96.20%
23	11719	G	A	156	98.10%
24	12280	A	G	138	98.60%
25	12705	C	T	132	97.00%
26	13105	A	G	147	98.60%
27	13886	T	C	156	97.40%
28	14284	C	T	155	98.10%
29	14287	T	C	156	96.20%
30	14634	T	C	156	95.50%
31	14766	C	T	148	99.30%
32	15110	G	A	143	98.60%
33	15301	G	A	148	97.30%
34	15326	A	G	155	97.40%
35	16124	T	C	145	97.90%
36	16223	C	T	155	98.10%
37	16519	T	C	122	95.90%

Table S7. RC27 vs O395 genomic island comparison

Class	GI§	Chr	Description§	O395 position	Pairwise identity	Confirmed?
Post "PG-2"§	GI-11	sm	Kappa prophage	1012862-1045566	99.9%	Yes
	GI-14	sm	Hypothetical proteins	814223-832392	99.9%	Yes
	GI-21	lg	Mu-like prophage	689569-722653	100%	Yes
		lg	Mu-like prophage	766615-799585	n/a	*Unknown*
	GI-23	lg	Putative prophage	2499645-2522558	100%	Yes
	GI-24	lg	Putative prophage (CRISPR-associated proteins)	2825203-2840525	91.1%	Yes

§ - From reference 34

Table S8. Results of the integron region core genome and O395-specific genes

Cluster§	Locus tag§	Description*	O395 position	%GC	Avg unique coverage	% reference covered @1X	Confirmed?
		Full "superintegron"					
-	VC0395_0740 to VC0395_0938	PA1849 is missing GI-14 (814,223-832,392)	799827-916350	41.0%	23.3	80.2%	Y
		Core genome integron genes§					
1	VC0395_0867	hypothetical protein	876486-876668	41.8%	50.4	100.0%	Y
1	VC0395_0907	hypothetical protein	898632-898814	44.8%	43.0	100.0%	Y
1	VC0395_0884	hypothetical protein	884967-885149	44.8%	42.1	100.0%	Y
2	VC0395_0802	putative acetyltransferase	838538-838972	41.7%	12.6	98.2%	Y
3	VC0395_0913	putative lipoprotein	900754-901179	41.6%	24.9	100.0%	Y
3	VC0395_0840	putative lipoprotein	863244-863669	40.9%	20.6	100.0%	Y
3	VC0395_0901	hypothetical protein	894878-895303	40.1%	21.2	100.0%	Y
3	VC0395_0819	hypothetical protein	850309-850704	40.1%	20.5	100.0%	Y
4	VC0395_0832	hypothetical protein	858808-858987	44.9%	71.9	100.0%	Y
4	VC0395_0902	hypothetical protein	895487-895666	44.3%	41.1	100.0%	Y
4	VC0395_0857	hypothetical protein	872107-872289	44.3%	48.9	100.0%	Y
5	VC0395_0833	hypothetical protein	858972-859106	43.4%	24.2	100.0%	Y
5	VC0395_0910	hypothetical protein	899789-899929	40.8%	27.8	100.0%	Y
7	VC0395_0938	intI4	915388-916350	41.6%	18.0	100.0%	Y
8	VC0395_0742	CopG family transcriptional regulator	801567-801935	44.5%	19.9	100.0%	Y
9	VC0395_0831	putative acetyltransferase	858099-858614	39.0%	18.2	97.7%	Y
9	VC0395_0741	putative acetyltransferase	800942-801535	39.9%	31.8	100.0%	Y
		O395-specific genes§					
308	VC0395_0931	hypothetical protein	910606-910719	41.9%	56.5	100.0%	Y
309	VC0395_0845	hypothetical protein	864958-865884	38.3%	4.8	83.5%	**Y?**
310	VC0395_0783	GI-14; hypothetical protein	826832-826984	32.7%	0.1	13.1%	**N**
311	VC0395_0885	hypothetical protein	885210-885380	43.0%	47.7	100.0%	Y
312	VC0395_0745	hypothetical protein	803209-803358	39.1%	51.3	100.0%	Y
313	VC0395_0876	hypothetical protein	881858-881959	54.6%	91.2	100.0%	Y
314	VC0395_0756	hypothetical protein	809326-809460	40.4%	50	100.0%	Y
315	VC0395_0838	hypothetical protein	862045-862185	3450.0%	8.1	100.0%	Y
316	VC0395_0780	GI-14; hypothetical protein	825158-825496	35.8%	0.3	16.2%	**N**
317	VC0395_0777	GI-14; hypothetical protein	824612-824764	40.5%	0	0.0%	**N**
318	VC0395_0909	hypothetical protein	899193-899369	42.8%	21.3	100.0%	Y
319	VC0395_0779	GI-14; hypothetical protein	825025-825141	35.0%	0	0.0%	**N**
320	VC0395_0859	hypothetical protein	872584-873033	34.8%	5.3	74.0%	**Y?**

321	VC0395_0774	GI-14; hypothetical protein	822322-822504	35.3%	0.2	10.4%	**N**
322	VC0395_0843	hypothetical protein	864432-864818	44.9%	16.2	100.0%	Y
323	VC0395_0789	GI-14; hypothetical protein	830312-830647	34.5%	0	0.0%	**N**
324	VC0395_0782	GI-14; hypothetical protein	826271-826747	36.6%	0.1	8.4%	**N**
325	VC0395_0912	hypothetical protein	900502-900687	43.9%	75.9	100.0%	Y
326	VC0395_0863	hypothetical protein	874509-874637	52.6%	89.4	100.0%	Y
327	VC0395_0824	blc-4	853496-854011	41.5%	14.7	100.0%	Y
328	VC0395_0908	hypothetical protein	898875-899030	42.9%	15.6	100.0%	Y
329	VC0395_0933	hypothetical protein	912303-912494	44.4%	69.9	100.0%	Y
330	VC0395_0847	hypothetical protein	866735-866941	41.8%	45.1	100.0%	Y
331	VC0395_0770	GI-14; hypothetical protein	817596-817973	41.0%	0.1	5.3%	**N**
332	VC0395_0860	isochorismatase family protein	873168-873707	38.2%	6.9	94.1%	**Y?**
333	VC0395_0788	GI-14; hypothetical protein	829454-830008	37.4%	0.7	25.0%	**N**
334	VC0395_0882	hypothetical protein	884165-884278	40.2%	59.8	100.0%	Y
335	VC0395_0791	GI-14 adjacent; hypothetical protein	832392-832850	39.7%	0.1	9.2%	**N**
336	VC0395_0852	hypothetical protein	868887-869006	45.8%	51.5	100.0%	Y
337	VC0395_0878	blc-4	882450-882695	41.3%	19.7	100.0%	Y
338	VC0395_0825	hypothetical protein	853923-854087	41.3%	58.5	100.0%	Y
339	VC0395_0879	hypothetical protein	882877-883041	41.8%	53.5	100.0%	Y
340	VC0395_0773	GI-14; hypothetical protein	822079-822315	36.7%	0.2	10.5%	**N**
341	VC0395_0841	hypothetical protein	863769-863882	36.8	43.4	100.0%	Y
342	VC0395_0856	hypothetical protein	871207-871899	34.1%	3.1	58.0%	**Y?**
343	VC0395_0781	GI-14; hypothetical protein	825518-825817	38.7%	0	0.0%	**N**
344	VC0395_0854	putative acetyltransferase	869432-869545	38.5%	66.2	100.0%	Y
345	VC0395_0767	GI-14; hypothetical protein	814797-814934	40.6%	0	0.0%	**N**
346	VC0395_0769	GI-14; hypothetical protein	817180-817575	36.4%	0.4	11.9%	**N**
347	VC0395_0837	hypothetical protein	861740-861930	41.7%	79.8	100.0%	Y
348	VC0395_0771	GI-14; hypothetical protein	817990-818265	35.0%	0.1	14.5%	**N**
349	VC0395_0839	hypothetical protein	862212-862469	38.2%	11.9	89.1%	Y
350	VC0395_0846	hypothetical protein	866032-866395	38.3%	8.4	96.1%	Y
351	VC0395_0848	hypothetical protein	867226-867807	37.0%	5.4	89.3%	**Y?**
352	VC0395_0924	hypothetical protein	906739-906888	44.1%	65.2	100.0%	Y

§ - From reference 18

*** -** Gene names, locus ID, and locus descriptions taken from NC_009456 chromosomal annotations (http://www.ncbi.nlm.nih.gov/bioproject?term=PRJNA58425).

VI. SUPPLEMENTARY REFERENCES

1. Shippen E. Memoir of John Neill, M.D., Late Emeritus Professor of Clinical Surgery in the University of Pennsylvania [Read October 6, 1880.]. In: Ashhurst Jr. J, ed. Transactions of the College of Physicians of Philadelphia. Philadelphia, PA: Lindsay & Blakiston; 1881:cxli-clvi.
2. Okello JBA, Zurek J, Devault AM, et al. Comparison of methods in the recovery of nucleic acids from archival formalin-fixed paraffin-embedded autopsy tissues. Anal Biochem 2010;400:110-7.
3. Meyer M, Kircher M. Illumina sequencing library preparation for highly multiplexed target capture and sequencing. Cold Spring Harbor Protoc 2010;2010:1-10, 1-7.
4. Kircher M, Sawyer S, Meyer M. Double indexing overcomes inaccuracies in multiplex sequencing on the Illumina platform. Nucleic Acids Res 2012;40.
5. Bos KI. Doctoral Thesis: Genetic Investigations into the Black Death. Hamilton, ON, Canada: McMaster University; 2011.
6. Hodges E, Xuan Z, Balija V, et al. Genome-wide in situ exon capture for selective resequencing. Nat Genet 2007;39:1522-7.
7. Hodges E, Rooks M, Xuan ZY, et al. Hybrid selection of discrete genomic intervals on custom-designed microarrays for massively parallel sequencing. Nat Protoc 2009;4:960-74.
8. Agilent SureSelect DNA Capture Array Protocol. Version 1.0 ed: Agilent Technologies; 2009.
9. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnetjournal 2011;17.
10. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 2009;25:1754-60.
11. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics 2009;25:2078-9.
12. Ginolhac A, Rasmussen M, Gilbert MTP, Willerslev E, Orlando L. mapDamage: testing for damage patterns in ancient DNA sequences. Bioinformatics 2011;27:2153-5.
13. Talavera G, Castresana J. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 2007;56:564-77.
14. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Syst Biol 2010;59:307-21.
15. Drummond AJ, Suchard MA, Xie D, Rambaut A. Bayesian Phylogenetics with BEAUti and the BEAST 1.7. Mol Biol Evol 2012;29:1969-73.
16. Martin DP, Lemey P, Lott M, Moulton V, Posada D, Lefeuvre P. RDP3: a flexible and fast computer program for analyzing recombination. Bioinformatics 2010;26:2462-3.
17. Lee J, Choi S, Jeon Y-S, et al. Classification of hybrid and altered Vibrio cholerae strains by CTX prophage and RS1 element structure. J Microbiol 2009;47:783-8.
18. Marin MA, Vicente ACP. Architecture of the superintegron in Vibrio cholerae: identification of core and unique genes [v1; ref status: approved 1, http:/f1000r.es/w6]. F1000Research 2013;2.
19. Seed KD, Lazinski DW, Calderwood SB, Camilli A. A bacteriophage encodes its own CRISPR/Cas adaptive response to evade host innate immunity. Nature 2013;494:489-91.
20. Makarova KS, Haft DH, Barrangou R, et al. Evolution and classification of the CRISPR–Cas systems. Nat Rev Micro 2011;9:467-77.
21. Mekalanos JJ. The evolution of Vibrio cholerae as a Pathogen: Humana Press Inc, 999 Riverview Dr, Ste 208, Totowa, Nj 07512-1165 USA; 2011.
22. Hochhut B, Waldor MK. Site-specific integration of the conjugal Vibrio cholerae SXT element into prfC. Molecular Microbiology 1999;32:99-110.

23. Lin W, Fullner KJ, Clayton R, et al. Identification of a Vibrio cholerae RTX toxin gene cluster that is tightly linked to the cholera toxin prophage. Proc Natl Acad Sci U S A 1999;96:1071-6.
24. Karaolis DKR, Lan RT, Kaper JB, Reeves PR. Comparison of Vibrio cholerae pathogenicity islands in sixth and seventh pandemic strains. Infect Immun 2001;69:1947-52.
25. Jermyn WS, Boyd EF. Characterization of a novel Vibrio pathogenicity island (VPI-2) encoding neuraminidase (nanH) among toxigenic Vibrio cholerae isolates. Microbiology 2002;148:3681-93.
26. Waldor MK, Mekalanos JJ. Lysogenic conversion by a filamentous phage encoding cholera toxin. Science 1996;272:1910-4.
27. Dziejman M, Balon E, Boyd D, Fraser CM, Heidelberg JF, Mekalanos JJ. Comparative genomic analysis of Vibrio cholerae: Genes that correlate with cholera endemic and pandemic disease. Proc Natl Acad Sci U S A 2002;99:1556-61.
28. Davis BM, Kimsey HH, Kane AV, Waldor MK. A satellite phage-encoded antirepressor induces repressor aggregation and cholera toxin gene transfer. Embo J 2002;21:4240-9.
29. Yamasaki S, Shimizu T, Hoshino K, et al. The genes responsible for O-antigen synthesis of Vibrio cholerae O139 are closely related to those of Vibrio cholerae O22. Gene 1999;237:321-32.
30. Andrews RM, Kubacka I, Chinnery PF, Lightowlers RN, Turnbull DM, Howell N. Reanalysis and revision of the Cambridge reference sequence for human mitochondrial DNA. Nat Genet 1999;23:147-.
31. Feng L, Reeves PR, Lan R, et al. A recalibrated molecular clock and independent origins for the cholera pandemic clones. PLoS ONE 2008;3:e4053.
32. Mutreja A, Kim DW, Thomson NR, et al. Evidence for several waves of global transmission in the seventh cholera pandemic. Nature 2011;477:462-U111.
33. Heidelberg JF, Eisen JA, Nelson WC, et al. DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. Nature 2000;406:477-83.
34. Chun J, Grim CJ, Hasan NA, et al. Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic Vibrio cholerae. Proc Natl Acad Sci U S A 2009;106:15442-7.
35. de Sa Morais LLC, Garza DR, Loureiro ECB, et al. Complete genome sequence of a sucrose-nonfermenting epidemic strain of Vibrio cholerae O1 from Brazil. J Bacteriol 2012;194:2772.
36. Reimer AR, Van Domselaar G, Stroika S, et al. Comparative genomics of Vibrio cholerae from Haiti, Asia, and Africa. Emerg Infect Dis 2011;17:2113-21.
37. Aydanian A, Tang L, Morris JG, Johnson JA, Stine OC. Genetic Diversity of O-Antigen Biosynthesis Regions in Vibrio cholerae. Appl Environ Microbiol 2011;77:2247-53.