LETTER

doi:10.1038/nature10392

Evidence for several waves of global transmission in the seventh cholera pandemic

Ankur Mutreja[1]*, Dong Wook Kim[2,3]*, Nicholas R. Thomson[1]*, Thomas R. Connor[1], Je Hee Lee[2,4], Samuel Kariuki[5], Nicholas J. Croucher[1], Seon Young Choi[2,4], Simon R. Harris[1], Michael Lebens[6], Swapan Kumar Niyogi[7], Eun Jin Kim[2], T. Ramamurthy[7], Jongsik Chun[4], James L. N. Wood[8], John D. Clemens[2], Cecil Czerkinsky[2], G. Balakrish Nair[7], Jan Holmgren[6], Julian Parkhill[1] & Gordon Dougan[1]

***Vibrio cholerae* is a globally important pathogen that is endemic in many areas of the world and causes 3–5 million reported cases of cholera every year. Historically, there have been seven acknowledged cholera pandemics; recent outbreaks in Zimbabwe and Haiti are included in the seventh and ongoing pandemic[1]. Only isolates in serogroup O1 (consisting of two biotypes known as 'classical' and 'El Tor') and the derivative O139 (refs 2, 3) can cause epidemic cholera[2]. It is believed that the first six cholera pandemics were caused by the classical biotype, but El Tor has subsequently spread globally and replaced the classical biotype in the current pandemic[1]. Detailed molecular epidemiological mapping of cholera has been compromised by a reliance on sub-genomic regions such as mobile elements to infer relationships, making El Tor isolates associated with the seventh pandemic seem superficially diverse. To understand the underlying phylogeny of the lineage responsible for the current pandemic, we identified high-resolution markers (single nucleotide polymorphisms; SNPs) in 154 whole-genome sequences of globally and temporally representative *V. cholerae* isolates. Using this phylogeny, we show here that the seventh pandemic has spread from the Bay of Bengal in at least three independent but overlapping waves with a common ancestor in the 1950s, and identify several transcontinental transmission events. Additionally, we show how the acquisition of the SXT family of antibiotic resistance elements has shaped pandemic spread, and show that this family was first acquired at least ten years before its discovery in *V. cholerae*.**

Whole-genome analysis is perhaps the ultimate approach to building a robust phylogeny in recently emerged pathogens, through the identification of SNPs and other rare genetic variants[4]. Therefore, we sequenced the genomes of 136 isolates of *V. cholerae*, the causative agent of several million cholera cases each year (http://www.who.int/mediacentre/factsheets/fs107/en/). These sequences, including 113 isolates from the seventh pandemic, were added to 18 previously published genomes[1,2,5] to produce a global genomic database from isolates collected in the course of a century. We included representative El Tor isolates collected in the past four decades and compared these to previously reported and novel genome sequences of both classical and non-O1 types[1,2].

The sequence reads were mapped to the reference sequence of El Tor N16961 (ref. 6), a seventh-pandemic *V. cholerae* that was isolated in Bangladesh in 1975 (see footnote to Supplementary Table 1) and the resulting consensus tree identified eight distinct phyletic lineages (L1–L8, see Supplementary Fig. 1 and Supplementary Table 1 for strain and lineage information), six of which incorporated O1 clinical isolates. The classical isolates formed a distinct, highly clustered group (L1), distant from the El Tor isolates of the seventh pandemic (L2). It is clear from Supplementary Fig. 1 that the classical and El Tor clades did not originate from a recent common ancestor and instead seem to be independent derivatives with distinct phylogenetic histories, consistent with previous proposals[2]. Isolates of L4 share a common ancestor with previously reported non-conventional O1 isolates[2] (Supplementary Fig. 2), and are likely to have acquired the O1 antigen genes by a recombination event onto a genetically distinct genome backbone. Isolates of L7 also have a distinct backbone, whereas L2, L3 (USA gulf coast strains), L5, L6 and L8 share a more 'El-Tor-like' genome backbone, and the L1 backbone is of the 'classical' type.

Genome-wide SNP analysis showed that the 123 El Tor isolates in the L2 cluster (Supplementary Fig. 1) differed from the reference by only 50–250 SNPs. With this large sample size we were able to construct a high-resolution phylogeny that shows unequivocally that the current pandemic is monophyletic and originated from a single source, providing a framework for future epidemiological and phenotypic analysis of *V. cholerae*, including transmission-tracking and typing.

Predicted recombined regions were identified, and along with genomic islands and mobile genetic elements, these were initially excluded from the phylogenetic analysis of seventh-pandemic isolates, to determine the underlying phylogeny. Notably, analysis of the tree (Fig. 1; see Supplementary Fig. 3 for a tree with strain names) provides clear evidence of a clonal expansion of the lineage, with a strong temporal signature. This is most clearly illustrated by the fact that the most divergent isolates from the N16961 reference are represented by the oldest seventh-pandemic isolate in our collection, A6, collected in 1957, together with the most recent Haitian isolates[5] from late 2010. We performed a linear regression analysis on all the L2 isolates to calculate the rate of SNP accumulation on the basis of the date of isolation and the root-to-tip distance. The shape of the tree and temporal signatures in Fig. 1 show a very consistent rate of SNP accumulation, 3.3 SNPs year^{-1} ($R^2 = 0.73$, Supplementary Fig. 4) in the core genome, emphasizing the tree's robustness and utility for transmission studies. The only exception to this is *V. cholerae* A4, a repeatedly passaged laboratory strain that was originally isolated in 1973 (Supplementary Figs 3 and 4). The estimated rate of mutation for our seventh-pandemic *V. cholerae* collection was 8.3×10^{-7} SNPs site^{-1} year^{-1}: between 5 and 2.5 times slower than the rate estimated for recent clonal expansions of some other human-pathogenic bacteria[4,7].

The seventh-pandemic tree can be subdivided into three major groups or clades by clustering using Bayesian analysis of population structure[8,9] (shown as waves 1–3 in Fig. 1); this clustering is mostly consistent with the cholera toxin (CTX) type of the three clades, which represent independent waves of transmission. Although examples of

[1]Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [2]International Vaccine Institute, SNU Research Park, Bongchun 7 dong, Kwanak, Seoul 151-919, Korea. [3]Department of Pharmacy, College of Pharmacy, Hanyang University, Kyeonggi-do 426-791, Korea. [4]Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul 151-742, Korea. [5]Centre for Microbiology Research, KEMRI at Kenyatta Hosp Compound, Off Ngong Road, PO Box 43640-00100, Kenya. [6]Department of Microbiology and Immunology and University of Gothenburg Vaccine Research Institute, The Sahlgrenska Academy at the University of Gothenburg, Box 435, 40530 Göteborg, Sweden. [7]National Institute of Cholera and Enteric Diseases, P-33, CIT Scheme XM, Beliaghata, Kolkata 700 010, India. [8]University of Cambridge, Department of Veterinary Medicine, Madingley Road, Cambridge CB3 0ES, UK.
*These authors contributed equally to this work.

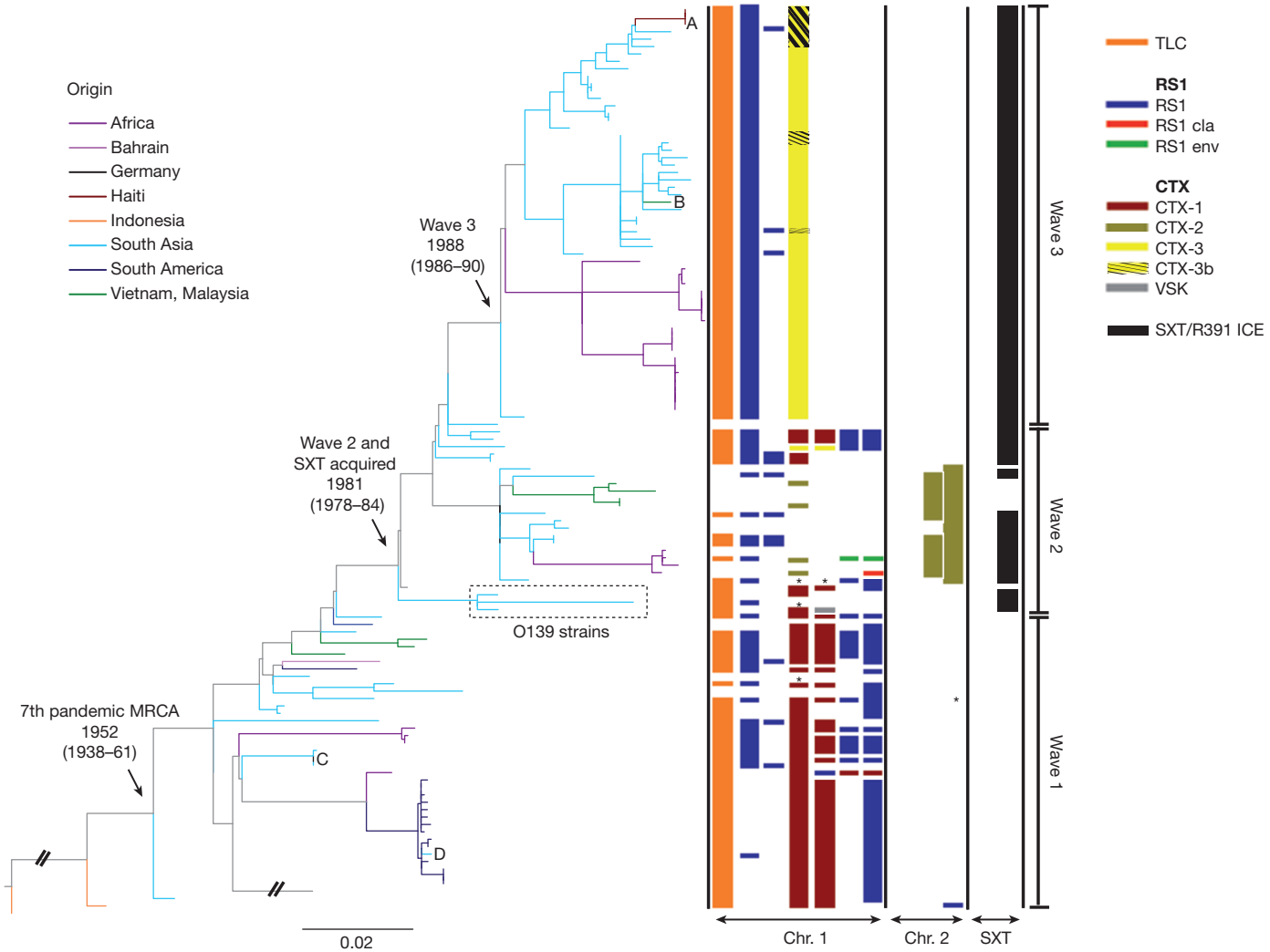

Figure 1 | A maximum-likelihood phylogenetic tree of the seventh pandemic lineage of *V. cholerae* based on SNP differences across the whole core genome, excluding probable recombination events. The pre-seventh-pandemic isolate M66 was used as an outgroup to root the tree. Branches are coloured on the basis of the region of isolation of the strains. The branches representing the three major waves are indicated on the far right. The nodes representing the MRCAs of the seventh pandemic, and subsequent waves 2 and 3, are indicated with arrows and labelled with inferred dates. The presence and type of CTX and SXT elements in each strain are shown to the right of the tree. The presence of toxin-linked cryptic (TLC) and repeated sequence 1 (RS1) elements is shown, but their number and position, respectively, are arbitrarily assigned. Cases of sporadic intercontinental transmission are marked A–D. The dates shown are the median estimates for the indicated nodes, taken from the results of the BEAST analysis. The scale is given as the number of substitutions per variable site; asterisks indicate that no data were available.

genetic determinants differentiating these three CTX types have previously been published[10], they have not been put into a phylogenetic context, undermining efforts to investigate the evolutionary aspects of their emergence. Perhaps as a result, there has been substantial uncertainty in naming new CTX types as they have been discovered. Our data shows that the first CTX type is canonical CTX El Tor and we propose that it is renamed CTX-1; for the other two we propose a new expandable nomenclature and class them as CTX-2 and CTX-3 (Supplementary Table 2).

Isolates spanning A18 to PRL5 (the lower clade in Fig. 1) represent wave 1, covering about 16 years (1977–1992). All isolates in this group lack the integrative and conjugative element (ICE) of the SXT/R391 family, encoding resistance to several antibiotics[11,12]. It is within this time period that seventh-pandemic cholera occurred in South America[6]. Our data show that the South American isolates form a discrete cluster, which also includes a single Angolan isolate collected in 1989. The position of the Angolan isolate at the base of the South American group indicates that transmission to South America may have been via Africa, as previously proposed[13]. We used BEAST[14] to translate evolutionary distance in SNPs into time (Supplementary Fig. 5) and this indicated that transmission to South America is likely to have occurred between 1981 and 1985. The branch harbouring this West African–South American (WASA) clade is distinguished from all other *V. cholerae* by the acquisition of novel VSP-2 genes[15] and a novel genomic island that we have denoted WASA1 (Supplementary Table 3). Notably, the Angolan isolate A5 and all the South American isolates are discriminated by just ten SNPs. Based on the accumulation rate of 3.3 SNPs year^{-1} (Supplementary Fig. 4), the 3-year time period between the isolation of A5 and the oldest South American isolate included in this study, A32, is consistent with previous studies indicating that cholera spread as a single epidemic[13].

The first acquisition of an SXT/R391 ICE lies at the point of transition from the wave-1 cluster to the wave-2 cluster. Using our dated phylogeny (Supplementary Fig. 5)[14], we were able to date this transition and the first acquisition of SXT/R391 ICE to 1978–84, ten years before its discovery in O139 strains, which also fits with the otherwise surprising discovery of SXT in a Vietnamese strain isolated before 1992 (ref. 16). This date would also correspond to the most recent common ancestor (MRCA) of the O1 and O139 serogroup isolates. Analysis of the diversity of the common regions of SXT/R391 ICEs in

our seventh-pandemic collection (Supplementary Fig. 6) shows that they are discriminated by 3,161 SNPs, compared to only 1,757 SNPs used to define the core whole-genome phylogeny in Fig. 1. This indicates either that there have been several recombination events within these ICEs, or that they have been acquired independently several times on the tree[11]. Isolates from wave 2 represent a discrete cluster that shows a complex pattern of accessory elements in the CTX locus (Fig. 1) and a wide phylogeographical distribution. It is also notable that isolates collected in Vietnam in 1995–2004 and strain A109 are the only wave-2 isolates studied from this time period that lack an SXT/R391 ICE. We examined the genomic locus in these clones that marks the point of insertion of SXT/R391 ICE in all other *V. cholerae* isolates and found no remnants of this conjugative element, which may have been lost from this lineage (no 'scar' in DNA sequence is expected after the precise excision of SXT/R391 ICE).

Ignoring the CTX-related genomic regions, the seventh-pandemic L2 isolates show relatively little evidence of recombination either within or from outside the tree. On the basis of the SNP distribution, 1,930 out of 2,027 SNPs (Supplementary Table 4) are congruent with the tree, leaving 97 homoplasies that could be due to selection or homologous recombination among the L2 isolates. Only 270 SNPs were predicted to be due to homologous recombination from outside the tree. The only two branches in which the SNP distribution indicated considerable recombination were those leading to the WASA cluster (Supplementary Fig. 7) and the O139 serogroup. Aside from the acquisitions of CTX and the SXT/R391 ICEs, we found evidence of gene flux affecting only 155 other genes (Supplementary Figs 8 and 9 and Supplementary Table 3).

Also represented in our collection are two isolates of serogroup O139, which are known to have arisen from a homologous replacement of their O-antigen determinant into an El Tor genomic backbone[2,3,13]. CTX types that are different from El Tor, classical, CTX-2 and CTX-3 have been reported for the O139 serogroup[17–20]; however, the phylogenetic position of the two strains included in this study shows that O139 was derived from O1 El Tor and therefore represents another distinct but spatially restricted wave from the common source.

We were also able to date the ancestor of the El Tor seventh-pandemic lineage, L2, as having existed in 1827–1936 (Supplementary Fig. 5), which is consistent with the predicted date of origin from the linear regression plot (1910, Supplementary Fig. 4). This also corresponds well with the date of isolation of the first El Tor biotype strain in 1905 (ref. 21).

It is apparent from Fig. 1 that *V. cholerae* wave 1, which spread globally, was later replaced by the more geographically restricted wave 2 and wave 3, a phenomenon supported by local clinical observations and phage analysis[10]. This also reflects the fact that *V. cholerae* epidemics since 2003–2010 have been restricted to Africa and south Asia. Notably, the rates of SNP accumulation calculated independently for wave 1, wave 3 and wave 2 (2.3, 2.6 and 3.5 SNPs year^{-1} respectively) are consistent with the rate calculated over the whole collection period (Supplementary Fig. 4).

The clonal clustering of L2 isolates, the constant rate of SNP accumulation and the temporal and geographical distribution support the concept that the seventh pandemic has spread by periodic radiation from a single source population located in the Bay of Bengal, followed by local evolution and ultimately local extinction in non-endemic areas. This is evidenced by the disappearance of wave-1 isolates, followed by the independent expansion of waves 2 and 3, both derived from the same original population, occurring within seven years of each other. These two waves are clearly distinguished from the first by the acquisition of SXT/R391 ICEs (Fig. 1). Plotting the intercontinental spread of each wave onto the world map (Fig. 2) clearly shows that the *V. cholerae* seventh pandemic is sourced from a single, restricted geographical location but has spread in overlapping waves. In these ancestral waves, there are at least four recent long-range transmission events (A–D in Fig. 1), in which isolates clearly share a common ancestor with recent strains at distant locations, indicating that such events are not uncommon. The most recent example of this is the Haitian outbreak, in which strains share a very recent common ancestor with south-Asian strains at the tip of wave 3. The number of SNP differences, even at whole-genome resolution, between the Haitian and the most closely related Indian and Bangladeshi strains is very low. This demonstrates that the Haitian strains must have come from south Asia, at most within the last six years. However, the limited discrimination means that it may prove challenging to make country-specific inferences as to the origins of the Haitian strains on the basis of DNA sequence alone. For such conclusions to be robust, great care must be taken in the selection of samples for analysis.

Despite clear evidence of sporadic long-range transmission events that are likely to be associated with direct human carriage, the overall pattern seen in our data is one of continued local evolution of *V. cholerae* in the

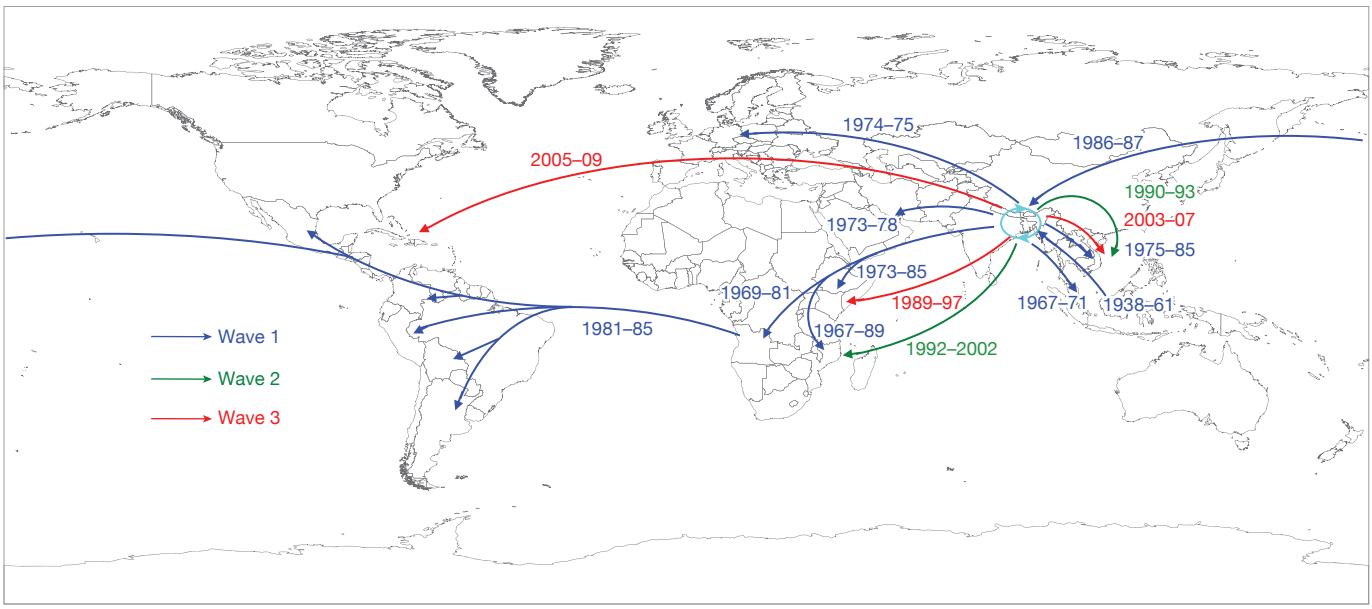

Figure 2 | Transmission events inferred for the seventh-pandemic phylogenetic tree, drawn on a global map. The date ranges shown for transmission events are taken from the BEAST analysis, and represent the median values for the MRCA of the transmitted strains (later bound), and the MRCA of the transmitted strains and their closest relative from the source location (earlier bound).

Bay of Bengal, with several independent waves of global transmission resulting in short-term epidemics in non-endemic countries. Although our sample set is substantial, there are clearly areas where geographical coverage is limited. However, the structure of the tree, with deep branches between the major waves, means that increasing the number of strains and the resolution further should only identify further independent waves of transmission. Indeed, we cannot rule out the possibility of an El Tor population persisting or evolving as a new wave of the seventh pandemic; for example, in areas such as China that were not sampled in this study.

One notable factor in the ongoing evolution of pandemic cholera was the acquisition of the SXT/R391-family antibiotic resistance element. The clinical use of the antibiotics tetracycline and furazolidone for cholera treatment started in 1963 and 1968 respectively, about 15 years before our prediction of the first acquisition of an SXT/R391 ICE (1978–1984). Our analysis provides a robust framework for elucidating the evolution of the seventh pandemic further, and for studying the local evolution, particularly in the Bay of Bengal, that has such a key role in the evolution of cholera.

METHODS SUMMARY

Genomic libraries were created for each sample, followed by multiplex sequencing on an Illumina GAIIx analyser. The 54-base paired-end reads obtained were mapped against N16961 El Tor as a reference and SNPs in the core genome were identified as described in Methods. The SNPs were used to draw a whole core-genome phylogeny as described in ref. 4. The final SNP alignment was used to perform BEAST[14] analysis and to confirm the output of linear regression analysis. The three cholera waves reported in the seventh-pandemic phylogeny were confirmed using BAPS[8,9]. The raw Illumina data were also assembled *de novo* (see Methods) so that pairwise genome comparisons could be made. A new and expandable nomenclature system describing the CTX trends seen in the last 40 years was proposed following the rationale described in Methods.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 3 June; accepted 26 July 2011.
Published online 24 August 2011.

1. Chin, C. S. *et al.* The origin of the Haitian cholera outbreak strain. *N. Engl. J. Med.* **364,** 33–42 (2011).
2. Chun, J. *et al.* Comparative genomics reveals mechanism for short-term and long-term clonal transitions in pandemic *Vibrio cholerae*. *Proc. Natl Acad. Sci. USA* **106,** 15442–15447 (2009).
3. Hochhut, B. & Waldor, M. K. Site-specific integration of the conjugal *Vibrio cholerae* SXT element into *prfC*. *Mol. Microbiol.* **32,** 99–110 (1999).
4. Harris, S. R. *et al.* Evolution of MRSA during hospital transmission and intercontinental spread. *Science* **327,** 469–474 (2010).
5. Update: cholera outbreak—Haiti, 2010. *MMWR Morb. Mortal Wkly Rep.* **59,** 1473–1479 (2010).
6. Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature* **406,** 477–483 (2000).
7. Croucher, N. J. *et al.* Rapid pneumococcal evolution in response to clinical interventions. *Science* **331,** 430–434 (2011).
8. Corander, J., Marttinen, P., Siren, J. & Tang, J. Enhanced Bayesian modelling in BAPS software for learning genetic structures of populations. *BMC Bioinformatics* **9,** 539 (2008).
9. Corander, J., Waldmann, P. & Sillanpaa, M. J. Bayesian analysis of genetic differentiation between populations. *Genetics* **163,** 367–374 (2003).
10. Safa, A., Nair, G. B. & Kong, R. Y. Evolution of new variants of *Vibrio cholerae* O1. *Trends Microbiol.* **18,** 46–54 (2010).
11. Garriss, G., Waldor, M. K. & Burrus, V. Mobile antibiotic resistance encoding elements promote their own diversity. *PLoS Genet.* **5,** e1000775 (2009).
12. Wozniak, R. A. *et al.* Comparative ICE genomics: insights into the evolution of the SXT/R391 family of ICEs. *PLoS Genet.* **5,** e1000786 (2009).
13. Lam, C., Octavia, S., Reeves, P., Wang, L. & Lan, R. Evolution of seventh cholera pandemic and origin of 1991 epidemic, Latin America. *Emerg. Infect. Dis.* **16,** 1130–1132 (2010).
14. Drummond, A. J., Ho, S. Y., Phillips, M. J. & Rambaut, A. Relaxed phylogenetics and dating with confidence. *PLoS Biol.* **4,** e88 (2006).
15. O'Shea, Y. A. *et al.* The *Vibrio* seventh pandemic island-II is a 26.9 kb genomic island present in *Vibrio cholerae* El Tor and O139 serogroup isolates that shows homology to a 43.4 kb genomic island in *V. vulnificus*. *Microbiology* **150,** 4053–4063 (2004).
16. Bani, S. *et al.* Molecular characterization of ICEVchVie0 and its disappearance in *Vibrio cholerae* O1 strains isolated in 2003 in Vietnam. *FEMS Microbiol. Lett.* **266,** 42–48 (2007).
17. Basu, A. *et al. Vibrio cholerae* O139 in Calcutta, 1992–1998: incidence, antibiograms, and genotypes. *Emerg. Infect. Dis.* **6,** 139–147 (2000).
18. Faruque, S. M. & Mekalanos, J. J. Pathogenicity islands and phages in *Vibrio cholerae* evolution. *Trends Microbiol.* **11,** 505–510 (2003).
19. Faruque, S. M. *et al.* The O139 serogroup of *Vibrio cholerae* comprises diverse clones of epidemic and nonepidemic strains derived from multiple *V. cholerae* O1 or non-O1 progenitors. *J. Infect. Dis.* **182,** 1161–1168 (2000).
20. Nair, G. B., Bhattacharya, S. K. & Deb, B. C. *Vibrio cholerae* O139 Bengal: the eighth pandemic strain of cholera. *Indian J. Public Health* **38,** 33–36 (1994).
21. Cvjetanovic, B. & Barua, D. The seventh pandemic of cholera. *Nature* **239,** 137–138 (1972).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements This work was supported by The Wellcome Trust grant 076964. The IVI is supported by the Governments of Korea, Sweden and Kuwait. D.W.K. was partially supported by grant RTI05-01-01 from the Ministry of Knowledge and Economy (MKE), Korea and by R01-2006-000-10255-0 from the Korea Science and Engineering Foundation; and J.L.N.W. was supported by the Alborada Trust and the RAPIDD program of the Science & Technology Directorate, Department of Homeland Security. Thanks to A. Camilli at Tufts University Medical School for providing the corrected N16961 sequence, to B.M. Nguyen at NIHE, Vietnam and M. Ansaruzzaman at ICDDR, Bangladesh for providing strains, and to M. Fookes at WTSI for training support.

Author Contributions A.M., D.W.K. and N.R.T. collected the data, analysed it and performed phylogenetic analyses and comparative genomics. J.H.L., S.Y.C., E.J.K. and J.C. analysed the CTX types. S.K., S.K.N. and T.R. were involved in strain collection and serogroup analysis. T.R.C. performed Bayesian analysis; N.J.C. and S.R.H. did the computational coding. J.L.N.W., J.D.C., C.C., G.B.K., J.H., N.R.T., J.P. and G.D. were involved in the study design. A.M., N.R.T., J.P., G.D., J.H., G.B.K., N.J.C., S.R.H., T.R.C., D.W.K. and M.L. contributed to the manuscript writing.

Author Information Reprints and permissions information is available at www.nature.com/reprints. All the genomic sequences have been submitted to the European Nucleotide Archive with the accession numbers listed in Supplementary Table 1. The authors declare no competing financial interests. Readers are welcome to comment on the online version of this article at www.nature.com/nature. Correspondence and requests for materials should be addressed to J.P. (parkhill@sanger.ac.uk).

METHODS

Genomic library creation and multiplex sequencing. Unique index-tagged libraries for each sample were created, and up to 12 separate libraries were sequenced in each of eight channels in Illumina Genome Analyser GAII cells with 54-base paired-end reads. The index-tag sequence information was used for downstream processing to assign reads to the individual samples[4].

Detection of SNPs in the core genome. The 54-base paired-end reads were mapped against the N16961 El Tor reference (accession numbers AE003852 and AE003853) and SNPs were identified as described in ref. 7. The unmapped reads and the sequences that were not present in all genomes were not considered a part of the core genome, and therefore SNPs from these regions were not included in the analysis. Appropriate SNP cutoffs were chosen to minimize the number of false-positive and false-negative calls; SNPs were filtered to remove those at sites with a SNP quality score lower than 30, and SNPs at sites with heterogeneous mappings were filtered out if the SNP was present in fewer than 75% of reads at that site. From the seventh-pandemic data set, high-density SNP clusters indicating possible recombination were excluded[7]. In total, 2,027 SNPs were detected in the core genome of the El Tor lineage. Of these, 270 SNPs were predicted to be due to recombination. Removing these provided a data set characterized by 1,757 SNPs: these were used to produce the final phylogeny.

Comparative genomics. Raw Illumina data were split to generate paired-end reads, and assembled using a *de novo* genome-assembly program, Velvet v0.7.03 (ref. 22), to generate a multi-contig draft genome for each of 133 *V. cholerae* strains[4]. The overlap parameters were optimized to give the highest N50 value. Because seventh-pandemic *V. cholerae* strains are closely related in the core, Abacas[23] was used to order the contigs using the N16961 El Tor strain as a reference, followed by annotation transfer from the reference strain to each draft genome[4]. Using the N16961 sequence as a database to perform a TBLASTX[24] for each draft genome, a genome comparison file was generated that was subsequently used in the Artemis comparison tool[25] to compare the genomes manually and search for novel genomic islands.

Phylogenetic analysis. A phylogeny was drawn for *V. cholerae* using RAxML v0.7.4 (ref. 26) to estimate the trees for all SNPs called from the core genome. The general time-reversible model with gamma correction was used for among-site rate variation for ten initial trees[4]. USA gulf coast strains A215 and A325, which have substantially different core genomes from all other strains in our collection, were used as an outgroup to root the global phylogeny (Supplementary Fig. 1), whereas a pre-seventh-pandemic strain, M66 (accession numbers CP001233 and CP001234), and strain A6 (from our collection), were used to root the seventh-pandemic phylogenetic tree (Fig. 1).

CTX prophage analysis. For each strain, the CTX structure and the sequence of *rst*A, *rst*R and *ctx*B was determined as in refs 27 and 28.

Linear regression and Bayesian analysis. The phylogram for the seventh pandemic was exported to Path-O-Gen v1.3 (http://tree.bio.ed.ac.uk/software/pathogen) and a linear regression plot for isolation date versus root-to-tip distance was generated. The same plot was also constructed individually for the three waves, but A4, being a laboratory strain, was excluded from the latter analysis.

The presence of three waves was checked, and their makeup was determined, using a BAPS analysis performed on the SNP alignment containing the unique SNP patterns from the seventh-pandemic isolates. The program was run using the BAPS individual mixture model and three independent iterations were performed using an upper limit for the number of populations of 20, 21 and 22 to obtain optimal partitioning of the sample. The dates for the acquisition of SXT and the ancestors of the three waves were inferred using the Bayesian Markov chain Monte Carlo framework BEAST[29]. We used the final SNP alignment with recombinant sites removed and fixed the tree topology to the phylogeny produced by RAxML, as described above. We used BEAST to estimate the rates of evolution on the branches of the tree using a relaxed molecular clock[14], which allows rates of evolution to vary amongst the branches of the tree. BEAST produced estimates for the dates of branching events on the tree by sampling dates of divergence between isolates from their joint posterior distribution, in which the sequences are constrained by their known date of isolation. The data were analysed using a coalescent constant population size and a general time-reversible model with gamma correction. The results were produced from three independent chains of 50 million steps each, sampled every 10,000 steps to ensure good mixing. The first 5 million steps of each chain were discarded as a burn-in. The results were combined using Log Combiner, and the maximum clade credibility tree was generated using Tree Annotator, both parts of the BEAST package (http://tree.bio.ed.ac.uk/software/beast/). Convergence and the effective sample-size values were checked using Tracer 1.5 (available from http://tree.bio.ed.ac.uk/software/tracer). ESS values in excess of 200 were obtained for all parameters.

Nomenclature. The seventh-pandemic cholera strains were clearly distinguished by three waves and we therefore propose their CTX types to be CTX-1, CTX-2 and CTX-3 under the new nomenclature scheme (see Supplementary Table 2). Our nomenclature system is expandable and would be suitable for naming any new seventh-pandemic *V. cholerae* strains. With CTX-1 representing canonical El Tor, we followed the rationale: (1) For CTX-1 to CTX-2, because there was a shift of $rstR^{\text{El Tor}}$ to $rstR^{\text{Classical}}$, $rstA^{\text{El Tor}}$ to $rstA^{\text{Classical + El Tor}}$ and $ctxB^{\text{El Tor}}$ to $ctxB^{\text{Classical}}$, we called it CTX-2; (2) for CTX-1 to CTX-3, because there was a shift of $ctxB^{\text{El Tor}}$ to $ctxB^{\text{Classical}}$, we called it CTX-3; (3) for CTX-3 to CTX-3b, because there was only one SNP mutation in $ctxB^{\text{Classical}}$ from CTX-2 and rest was identical, we called it the next variant of CTX-3, which is CTX-3b.

In summary, if there is a shift of any gene from one biotype to another, the new CTX will be called CTX-n: thus the next strains fitting these criteria will be called CTX-4. However, if there is a mutation(s) that does not lead to a shift of the gene to another biotype gene, CTX-1b, CTX-1c or CTX-2b; CTX-2c or CTX-3b; CTX-3c and so on should be followed as appropriate.

22. Zerbino, D. R. & Birney, E. Velvet: algorithms for *de novo* short read assembly using de Bruijn graphs. *Genome Res.* **18,** 821–829 (2008).
23. Assefa, S., Keane, T. M., Otto, T. D., Newbold, C. & Berriman, M. ABACAS: algorithm-based automatic contiguation of assembled sequences. *Bioinformatics* **25,** 1968–1969 (2009).
24. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215,** 403–410 (1990).
25. Carver, T. *et al.* Artemis and ACT: viewing, annotating and comparing sequences stored in a relational database. *Bioinformatics* **24,** 2672–2676 (2008).
26. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
27. Lee, J. H. *et al.* Classification of hybrid and altered *Vibrio cholerae* strains by CTX prophage and RS1 element structure. *J. Microbiol.* **47,** 783–788 (2009).
28. Nguyen, B. M. *et al.* Cholera outbreaks caused by an altered *Vibrio cholerae* O1 El Tor biotype strain producing classical cholera toxin B in Vietnam in 2007 to 2008. *J. Clin. Microbiol.* **47,** 1568–1571 (2009).
29. Drummond, A. J. & Rambaut, A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* **7,** 214 (2007).