The Origin of the Haitian Cholera Outbreak Strain

Chen-Shan Chin, Ph.D., Jon Sorenson, Ph.D., Jason B. Harris, M.D., William P. Robins, Ph.D., Richelle C. Charles, M.D., Roger R. Jean-Charles, M.D., James Bullard, Ph.D., Dale R. Webster, Ph.D., Andrew Kasarskis, Ph.D., Paul Peluso, Ph.D., Ellen E. Paxinos, Ph.D., Yoshitoku Yamaichi, Ph.D., Stephen B. Calderwood, M.D., John J. Mekalanos, Ph.D., Eric E. Schadt, Ph.D., and Matthew K. Waldor, M.D., Ph.D.

BACKGROUND
Although cholera has been present in Latin America since 1991, it had not been epidemic in Haiti for at least 100 years. Recently, however, there has been a severe outbreak of cholera in Haiti.

METHODS
We used third-generation single-molecule real-time DNA sequencing to determine the genome sequences of 2 clinical Vibrio cholerae isolates from the current outbreak in Haiti, 1 strain that caused cholera in Latin America in 1991, and 2 strains isolated in South Asia in 2002 and 2008. Using primary sequence data, we compared the genomes of these 5 strains and a set of previously obtained partial genomic sequences of 23 diverse strains of V. cholerae to assess the likely origin of the cholera outbreak in Haiti.

RESULTS
Both single-nucleotide variations and the presence and structure of hypervariable chromosomal elements indicate that there is a close relationship between the Haitian isolates and variant V. cholerae El Tor O1 strains isolated in Bangladesh in 2002 and 2008. In contrast, analysis of genomic variation of the Haitian isolates reveals a more distant relationship with circulating South American isolates.

CONCLUSIONS
The Haitian epidemic is probably the result of the introduction, through human activity, of a V. cholerae strain from a distant geographic source. (Funded by the National Institute of Allergy and Infectious Diseases and the Howard Hughes Medical Institute.)
The outbreak of cholera that began in Haiti in late October 2010 illustrates the continued public health threat of this ancient scourge. Cholera, an acutely dehydrating diarrheal disease that can rapidly kill its victims, is caused by Vibrio cholerae, a gram-negative bacterium. This disease, which is usually transmitted through contaminated water, can and has spread in an explosive fashion. In the weeks since cases were first confirmed in the Artibonite province of Haiti on October 19, 2010, the disease has reached all 10 provinces in Haiti and has spread to the neighboring Dominican Republic on the island of Hispaniola. Of the more than 93,000 persons who have been sickened from the outbreak, more than 2100 have died, according to the Haitian Ministry of Public Health and Population (www.mspp.gouv.ht/site/index.php), and it is thought that the epidemic has not yet peaked. Cholera epidemics had not been reported in Haiti for more than a century, and the origin of the Haitian V. cholerae outbreak has been the subject of some controversy.

Traditionally, V. cholerae strains are classified into serogroups on the basis of the structure of an outer-membrane O antigen and into biotypes on the basis of a variety of biochemical and microbiologic tests. The ongoing seventh pandemic of cholera is caused by the V. cholerae El Tor biotype of serogroup O1 (El Tor O1), which has replaced the previous “classical” biotype and has spread globally since its appearance in Indonesia in 1961. It reached the Americas in 1991, beginning in Peru and then spreading throughout much of South America and Central America, where it has since become endemic; however, the strains of V. cholerae El Tor O1 that are now endemic in South America and Central America had not previously been reported to have caused cholera on Hispaniola. Analyses carried out by Haitian and U.S. laboratories have indicated that the current outbreak strain in Haiti is also V. cholerae El Tor O1 and thus is related to strains that are causing the ongoing seventh pandemic of cholera.

Both genetic and phenotypic diversity have arisen among circulating strains of V. cholerae El Tor O1, reflecting the acquisition, loss, or alteration of mobile genetic elements (for this and other key terms, see the Glossary), including CTX phage, which bears the genes encoding cholera toxin; genomic islands; and SXT-family integrative and conjugative elements, which often encode resistance to several antibiotics. Single-nucleotide variations (SNVs) and insertions and deletions have also been detected in the core V. cholerae genome. Such heterogeneity has been used to group strains and to model and understand their transmission around the globe and is most comprehensively captured by sequencing genomic DNA. Second-generation DNA-sequencing technologies, although greatly productive, require a week or more to generate DNA sequence at high coverage and produce reads that are much shorter than those produced with first-generation sequencing technologies — making it difficult to characterize DNA variation in repeat regions. Third-generation single-molecule real-time sequencing involves direct observation of the DNA polymerase while it synthesizes a strand of DNA; thus, it is much faster than previously developed methods and provides a comparatively long read length. We therefore used a third-generation, single-molecule, real-time DNA sequencing method to determine the genome sequences of two Haitian V. cholerae isolates and three additional V. cholerae clinical isolates from other regions of the world, allowing us to determine the probable origin of the cholera outbreak strain in Haiti.

**METHODS**

**PATIENTS AND SAMPLES**

Samples of spontaneously passed stool from two patients who had received a clinical diagnosis of cholera were cultured. Both patients received standard medical treatment for cholera, as appropriate to their clinical conditions. Bacterial isolates (H1 and H2) were shipped to the United States, with the use of an import license for this purpose (2010-10-108) that was provided through the Centers for Disease Control and Prevention (CDC). Isolates were identified as V. cholerae and were determined to be susceptible to tetracycline and erythromycin but resistant to trimethoprim–sulfamethoxazole and nalidixic acid. The use of bacterial isolates derived from discarded stool samples and that do not have individual patient identifiers is exempt from regulations regarding research on human subjects. Existing clinical isolates from the 1991 outbreak in Peru, strain C6706 (C6); the 2008 outbreak in Bangladesh, strain MDC126 (M4); and the 1971 outbreak in Bangla-
desh, strain N16961 (N5) were cultured as described in the Supplementary Appendix, available with the full text of this article at NEJM.org.

**DNA Preparation and Sequencing**

We isolated genomic DNA from each of the C6, N5, M4, H1, and H2 strains and sequenced it using previously described methods.\(^{15}\) More specifically, we constructed DNA libraries comprising SMRTbell constructs, each of which was bound to a DNA polymerase and sequenced in a manner similar to that described previously,\(^ {16}\) using the PacBio RS sequencing system (Pacific Biosciences). For additional details regarding DNA
sequencing, resequencing analysis, and detection of DNA variations, see the Supplementary Appendix. Methods for the reconstruction of phylogenetic trees and the characterization of VSP-2 (a genomic island), SXT, and the superintegron are provided in the Supplementary Appendix.

RESULTS

STRUCTURES OF FIVE V. CHOLERAE GENOMES

The H1 and H2 isolates were sequenced in less than 24 hours, with enough DNA sequencing reads generated in this time to cover the genomes 60 and 32 times, respectively. C6, M4, and N5 were similarly rapidly sequenced at coverages of 28, 37, and 36, respectively. (Table 1 in the Supplementary Appendix). We used previously obtained genome sequences of N16961,17 CIRS101,11 and MJ-123644 as reference genomes to facilitate genomewide characterization of the five sequenced isolates. When we mapped raw sequencing reads to the canonical N16961 reference, we identified copy-number variation — typically in hyper-recombinant genomic regions — affecting ribosomal RNAs, the V. cholerae superintegron, the SXT-integrative and conjugative element, and the seventh-pandemic genomic islands (VSP-1 and VSP-2).18 The five isolates showed a high degree of similarity, as well as notable structural variation (Fig. 1). The structures of the H1 and H2 genomes were identical (Fig. 1). The sequence from sample N5 matched the canonical reference strain from which it was purportedly cultured.

SNVS AND THE RELATEDNESS OF THE V. CHOLERAE STRAINS

A comparison of the SNVs of each strain also indicated that H1 and H2 were essentially identical and were more similar to the M4 strain from Asia than to the C6 strain from Peru or the canonical N16961 reference (Table 2 in the Supplementary Appendix). Although we used data from 20-times coverage to determine the SNVs present in each genome (GenBank accession number, SRP004712) for the comparative analyses, the key SNVs highlighted in Figure 1 were apparent after achieving 12-times coverage of the genomes of these isolates; we obtained 12-times coverage of the five genomes within 3 hours of sequencing.

In our initial assessment of the relatedness of the five sequenced isolates, we analyzed a set of 1588 conserved orthologous genes (encompassing approximately 1.8 Mb of DNA) that were previously reported to resolve the relatedness of different V. cholerae strains40 of diverse origin. We aligned the consensus sequences of those 1.8 Mb from C6, N5, M4, H1, and H2 with those of 23 previously sequenced V. cholerae strains40 and constructed a phylogenetic tree that unequivocally places H1 and H2 in the seventh-pandemic group. Although the Haitian strains are similar to isolates from Latin America (C6 from the 1991 outbreak in Peru and Africa (B33 from the 2004 outbreak in Mozambique), they are most closely related to recent South Asian isolates (M4 from the 2008 outbreak in Bangladesh and CIRS101 from the 2002 outbreak in Bangladesh) (Fig. 2A). H1 and H2 are only distantly related to the U.S. Gulf Coast isolates, such as strain 2740-80; the latter does not even cluster with seventh-pandemic strains (Fig. 2A).

Thirty SNVs have previously been shown to differentiate six groups within the seventh-pandemic strains.19 We compared the alleles of these SNVs from each of the five isolates with those from 78 cholera strains from the seventh pandemic and 3 cholera strains isolated before the seventh pandemic19 and constructed a phylogenetic tree (Fig. 2B). Six groups from the seventh pandemic are readily identified in this tree, with H1 and H2 falling into group V, which also includes variant strains from Bangladesh (CIRS101

Figure 1 (facing page). Sequence Depth of Coverage for the Five Vibrio cholerae Isolates.

The figure shows the observed sequencing depth of coverage for the five isolates we sequenced (H1, H2, M4, N5, and C6), relative to the published sequence from the two chromosomes of V. cholerae N16961. Areas in the genome in which the read coverage was more than 4 SD higher than the background coverage are plotted with green points (repetitive regions). Areas in the genome in which the read coverage was more than 4 SD lower than the background coverage are plotted with red points (missing segments). Regions in the outer ring show known strain markers that allow the typing of these five isolates with respect to each other and to published strains. The locations of discriminating markers of single-nucleotide variations (SNVs) are shown as yellow bands, and the locations of discriminating mobile elements are shown as orange bands. The identifiers outside the outermost circle correspond to the positions of known mobile elements10 and strain-specific SNV markers.19
and M4). The phylogeny highlights the distance between strains of group V and those of group II, the latter of which consists mainly of Latin American strains (including C6, the strain we sequenced) and African strains isolated between 1970 and 1998. It supports the conclusion that Haitian *V. cholerae* is more closely related to contemporary South Asian strains of *V. cholerae* than to Latin American strains. The placement of C6 in group II is consistent with a previously proposed hypothesis that Latin American strains of *V. cholerae* may have been introduced from Africa.20

**STRUCTURAL VARIATIONS IN THE SUPERINTEGRON, VSP-2, AND SXT**

Analyses of insertions and deletions in hyper-recombinant chromosomal elements, which are often mobile elements, can be used to complement the analysis of SNV markers in the establishment of the lineage of a given strain.10 We therefore assessed the sequences of 20 previously described hyper-recombinant chromosomal elements20 in the genomes of C6, N5, M4, H1, and H2 (Fig. 1 for the locations of these elements). The long read lengths that we obtained (the average read length of filtered H1 and H2 sequences was 954 bp, with 5% of the reads exceeding 2800 bases) are ideal for identifying structural variation, especially in the context of repeated DNA sequences. Of the 20 regions we examined, most were structurally conserved in the five strains we sequenced — consistent with the coverage results in Figure 1. However, we did observe structural variation in 3 of the 20 regions: superintegron, VSP-2, and SXT.

A map of the superintegron region from strains C6, N5, M4, and H1 is shown in Figure 3A. The superintegrons of C6 and N5 are structurally identical to that of the canonical reference strain N16961 (Table 1). In contrast, the superintegron structures of M4 and H1 are distinct from those of C6 and N5 (i.e., N16961); both M4 and H1 lack a segment that contains 41 open reading frames (Table 3 in the Supplementary Appendix). M4 is also missing a single open reading frame that is present in the H1 superintegron; otherwise, their genomic structures in this region are identical. Because the SNV data suggested that H1 (and H2) are more closely related to CIRS101 than to M4 (Fig. 2A), we also compared superintegron regions of the H1 and CIRS101 strains and found them to be structurally identical.

H1, M4, and C6 lack different overlapping segments of the VSP-2 region relative to N16961 (Table 3 and Fig. 1 in the Supplementary Appendix). The pattern of deletion in the VSP-2 sequence of CIRS101 is identical to that of H1, but not to that of M4, providing additional evidence that H1 is more closely related to CIRS101 than to M4 (Table 1).

SXT is a clinically important integrative and conjugative element that accounts for the dissemination of genes conferring resistance to several antibiotics in contemporary *V. cholerae* isolates.20 N16961 and Latin American epidemic strains (including C6706) are known to lack SXT and remain susceptible to antibiotics; not surprisingly, no reads from N5 or C6 mapped to a reference SXT sequence derived from the MJ-1236 strain (Table 1). However, structural analyses revealed that M4 and H1 contained very similar SXT elements and that both lack a closely related subset of the SXT genes that is present in MJ-1236 (Fig. 3B, and Table 3 in the Supplementary Appendix).

SNVs with biologic and epidemiologic significance have accumulated in the CTX prophage region. The gene encoding cholera toxin B subunit (ctxB) in isolate H1 (and H2) carries three nonsynonymous substitutions relative to N16961 (Fig. 3C). Two of these changes are characteristic of ctxB in classical strains of the sixth pandemic, and they have been detected in recent El Tor O1 strains (including CIRS101) from South Asia11 M4, H1, and H2 carry these two ctxB mutations. The third mutant allele, predicting the substitution of histidine with asparagine at position 20, (last line, Fig. 3C) has previously been observed only in El Tor variant strains from South Asia21 and in very recent isolates from West Africa.22

**RELATEDNESS OF H1 AND H2 TO OTHER ISOLATES OF HAITIAN V. CHOLERAE**

We compared the sequences of H1 and H2 to the unassembled genome sequence data of three in-
The origin of the cholera outbreak strain in Haiti is illustrated in the diagram above. The strain is a member of Group V, which is a cluster of strains that emerged in the 1990s. The strain is closely related to strains from Bangladesh and India, indicating that the outbreak may have originated from one of these regions. The specific strain, MJ-1236, was isolated from patients in Bangladesh in 1994 and is a representative of Group V.
dependently isolated Haitian strains that have been deposited by the CDC into the GenBank database (accession numbers, AELH00000000.1, AELI00000000.1, and AEIJ00000000.1). H1, H2, and the three isolates obtained by the CDC are virtually identical in all the regions previously shown to harbor structural variation. 10 The three coding mutations found in the ctxB gene of H1 and H2 are also present in each of the three CDC strains.

**Discussion**

The *V. cholerae* strain responsible for the expanding cholera epidemic in Haiti is nearly identical to so-called variant seventh-pandemic El Tor O1 strains that are predominant in South Asia, including Bangladesh. 23, 24 The shared ancestry of the Haitian epidemic strain and recent South Asian strains of *V. cholerae* is distinct from that of circulating Latin American and East African strains of *V. cholerae*. Patterns of DNA from Haitian strains and *V. cholerae* strains in a large collection held by the CDC, as determined by means of pulsed-field gel electrophoresis, also suggested that the Haitian strains of *V. cholerae* are most similar to re-
cent South Asian *V. cholerae* strains. Our comparative analysis of the H1 and H2 strains and three CDC isolates indicate that the Haitian cholera epidemic is clonal. Collectively, our data strongly suggest that the Haitian epidemic began with introduction of a *V. cholerae* strain into Haiti by human activity from a distant geographic source.

Our data distinguish the Haitian strains from those circulating in Latin America and the U.S. Gulf Coast and thus do not support the hypothesis that the Haitian strain arose from the local aquatic environment. It is therefore unlikely that climatic events led to the Haitian epidemic, as has been suggested in the case of other cholera epidemics. Understanding exactly how this South Asian variant strain of *V. cholerae* was introduced to Haiti will require further epidemiologic investigation.

The Haitian outbreak strains can be distinguished from earlier seventh-pandemic strains by several genetic polymorphisms, including those in *ctxB*. Alterations in the *ctxB* sequence in the context of other structural variations (e.g., within SXT and VSP-2) are hallmarks of the variant strains that have emerged in South Asia. Because these variant strains replaced previously dominant strains of the seventh pandemic in South Asia, it has been hypothesized that their unique genetic composition increases their relative fitness, perhaps as a consequence of increased pathogenicity. Specifically, by causing more severe dehydrating disease, variant strains increase their own dissemination through the increased production of infectious stools by their human hosts.

Our findings have policy implications for public health officials who are considering the deployment of vaccines or other measures for controlling cholera. The apparent introduction of cholera into Haiti through human activity emphasizes the concept that predicting outbreaks of infectious diseases requires a global rather than a local assessment of risk factors. The accidental introduction of South Asian variant *V. cholerae* El Tor into Haiti may have consequences beyond Haiti. The apparently higher relative fitness and increased antibiotic resistance of the South Asian strains and the ability of those strains to cause severe cholera suggest that the South Asian variant *V. cholerae* El Tor that is now in Haiti could displace the resident El Tor O1 seventh-pandemic strains in Latin America. It is likely that the Caribbean ecosystem may now be host to a set of genes, including classical biotype-like cholera toxin genes and the STX integrative and conjugative element, that were previously absent from this region. Clearly, the provision of adequate sanitation and clean water is essential for preventing the further spread of the Haitian cholera epidemic. Vaccination would also help to prevent the spread of disease, although cholera vaccines are in short supply. Our findings suggest that public health measures to counter the spread of cholera in Hispaniola

---

**Table 1. Gene Content of Hypervariable Elements in H1, H2, M4, and C6.**

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Has CTX</th>
<th>Missing ORFs in VSP-II region</th>
<th>Missing ORFs in Superintegron</th>
<th>Missing ORFs in SXT_MJ-1236</th>
<th>Missing ORFs in SXT_CIRS101</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sample</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>H1 and H2</td>
<td>Yes</td>
<td>Yes†</td>
<td>Yes</td>
<td>Yes†</td>
<td>No</td>
</tr>
<tr>
<td>M4</td>
<td>Yes</td>
<td>Yes†</td>
<td>Yes</td>
<td>Yes†</td>
<td>Yes</td>
</tr>
<tr>
<td>N5</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No hit‡</td>
<td>No hit‡</td>
</tr>
<tr>
<td>C6</td>
<td>Yes</td>
<td>Yes†</td>
<td>No</td>
<td>No hit‡</td>
<td>No hit‡</td>
</tr>
</tbody>
</table>

Reference genomes

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Has CTX</th>
<th>Missing ORFs in VSP-II region</th>
<th>Missing ORFs in Superintegron</th>
<th>Missing ORFs in SXT_MJ-1236</th>
<th>Missing ORFs in SXT_CIRS101</th>
</tr>
</thead>
<tbody>
<tr>
<td>CIRS101</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>ND</td>
<td>No</td>
</tr>
<tr>
<td>N16961</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>B33</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>MJ-1236</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No§</td>
<td>No§</td>
</tr>
</tbody>
</table>

* ND denotes not determined, and ORF open reading frame.† The missing regions are different in each of these.‡ No hit indicates that no SXT ORF was observed.§ Reference sequence comprises subsequences derived from MJ-1236.
could minimize the dissemination of the new South Asian strain and the virulence genes that it carries beyond the shores of this Caribbean island.

Work at Harvard Medical School and Brigham and Women’s Hospital was supported by grants from the National Institute of Allergy and Infectious Diseases (AI-018045, to Dr. Mekalanos; and AI R37-042347) and by a grant from the Howard Hughes Medical Institute to Dr. Waldor. Work at Massachusetts General Hospital was supported by a grant from the National Institute of Allergy and Infectious Diseases (AI058935, to Dr. Calderwood).

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

We thank the organizations and persons who continue to provide outstanding patient care in this outbreak and the Massachusetts General Hospital Office of Disaster Response/Center for Global Health and Project Hope, which allowed some of our team to assist as volunteers in this cholera outbreak; Steve Turner (Pacific Biosciences) for discussion and advice and Kristin Robertshaw for assistance in rendering Figure 2; Ali Bashar, Simon Chang, Janice Cheng, Pei-Lin Hsiung, Amruta Joshi, Dimitris Ilipolopoulos, Aaron Klammer, Deborah Kwo, Brianna LaMay, Steven Lin, Aseneth Lopez, Khai Luong, John Major, Patrick Marks, Phillip McClurg, Emilia Mollova, Huy Nguyen, Andy Pham, Ruben Pingu, Homero Rey, Robert Sebra, Marie Valdivosino, Susana Wang, and Jackie Yen at Pacific Biosciences for their assistance in rapidly preparing and sequencing the cholera samples and aiding in the analyses; and Brigid Davis, Wen Zheng, Dan Portnoy, and Steve Lory for discussion of our results and the manuscript.

REFERENCES