LA-UR- 09-01624

Approved for public release; distribution is unlimited.

Title:	Genome Assortment, Not Serogroup, Defines Vibrio cholerae Pandemic Strains
Author(s):	Jongsik Chun, Christopher J. Grim, Nur A. Hasan, Je Hee Lee, Seon Young Choi, Bradd J. Haley, Elisa Taviani, Yoon-Seong Jeon, Dong Wook Kim, Jae-Hak Lee, Thomas S. Brettin, David C. Bruce, Jean F. Challacombe, J. Chris Detter, Cliff S. Han, A. Christine Munk, Olga Chertkov, Linda Meincke, Elizabeth Saunders, Ronald A. Walters, Anwar Huq , G. Balakrish Nair and Rita R. Colwell
Intended for:	Nature

Los Alamos
NATIONAL LABORATORY
EST. 1943

Form 836 (7/06)

Genome Assortment, Not Serogroup, Defines *Vibrio cholerae* Pandemic Strains

Jongsik Chun [1,2,3], Christopher J. Grim [2], Nur A. Hasan [4,6], Je Hee Lee [1,3], Seon Young Choi [1,3], Bradd J. Haley[4], Elisa Taviani [4], Yoon-Seong Jeon [3], Dong Wook Kim [3], Jae-Hak Lee [1], Thomas S. Brettin [7], David C. Bruce [7], Jean F. Challacombe [7], J. Chris Detter [7], Cliff S. Han [7], A. Christine Munk [7], Olga Chertkov [7], Linda Meincke [7], Elizabeth Saunders [7], Ronald A. Walters[8], Anwar Huq [4], G. Balakrish Nair [9] and Rita R. Colwell [2,4,5]

[1], School of Biological Sciences and Institute of Microbiology, Seoul National University, Seoul 151-742, Republic of Korea,

[2,] Center of Bioinformatics and Computational Biology, University of Maryland Institute of Advanced Computer Studies, University of Maryland, College Park, MD 20742, U.S.A.

[3], International Vaccine Institute, Seoul, 151-818, Republic of Korea,

[4], Maryland Pathogen Research Institute, University of Maryland, College Park, MD 20742, USA

[5,] Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA

[6], International Center for Diarrheal Disease Research, Bangladesh, Dhaka-1000, Bangladesh

[7]DOE Joint Genome Institute, Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

[8], Pacific Northwest National Laboratory, Richland, WA 99352, USA

[9], National Institute of Cholera and Enteric Diseases, Beliaghata, Kolkata 700 010, India

Correspondence and requests for materials should be addressed to Rita R. Colwell (rcolwell@umiacs.umd.edu)

Vibrio cholerae, the causative agent of cholera, is a bacterium autochthonous to the aquatic environment, and a serious public health threat. _V. cholerae_ serogroup O1 is responsible for the previous two cholera pandemics, in which classical and El Tor biotypes were dominant in the 6th and the current 7th pandemics, respectively. Cholera researchers continually face newly emerging and re-emerging pathogenic clones carrying combinations of new serogroups as well as of phenotypic and genotypic properties. These genotype and phenotype changes have hampered control of the disease. Here we compare the complete genome sequences of 23 strains of _V. cholerae_ isolated from a variety of sources and geographical locations over the past 98 years in an effort to elucidate the evolutionary mechanisms governing genetic diversity and genesis of new pathogenic clones. The genome-based phylogeny revealed 12 distinct _V. cholerae_ phyletic lineages, of which one, designated the _V. cholerae_ core genome (CG), comprises both O1 classical and El Tor biotypes. All 7th pandemic clones share nearly identical gene content, i.e., the same genome backbone. The transition from 6th to 7th pandemic strains is defined here as a "shift" between pathogenic clones belonging to the same O1 serogroup, but from significantly different phyletic lineages within the CG clade. In contrast, transition among clones during the present 7th pandemic period can be characterized as a "drift" between clones, differentiated mainly by varying composition of laterally transferred genomic islands, resulting in emergence of variants,

exemplified by *V.cholerae* serogroup O139 and *V.cholerae* O1 El Tor hybrid clones that produce cholera toxin of classical biotype. Based on the comprehensive comparative genomics presented in this study it is concluded that *V. cholerae* undergoes extensive genetic recombination *via* lateral gene transfer, and, therefore, genome assortment, not serogroup, should be used to define pathogenic *V. cholerae* clones.

Vibrio cholerae, a bacterium autochthonous to the aquatic environment, is the causative agent of cholera, a severe, watery, life-threatening diarrheal disease. Historically, cholera bacteria have been serogrouped based on their somatic O antigens, with more than 200 serogroups identified to date [1]. Although strains from many of the serogroups of *V. cholerae* have caused either individual cases of mild gastroenteritis or local outbreaks of gastroenteritis, only the toxigenic strains of serogroups O1 and O139 have been identified as agents of cholera epidemics. Genes coding for cholera toxin, *ctx*AB, and other virulence factors have been shown to reside in bacteriophages and various mobile genetic elements. In addition, *V. cholerae* serogroup O1 is differentiated into two biotypes, classical and El Tor, by a combination of biochemical traits and sensitivity to specific bacteriophages [2].

Throughout human history cholera pandemics have been recorded with seven such pandemics characterized over the past hundred or more years. Today the disease remains endemic only in developing countries, even though *V. cholerae* is native to estuaries and river systems throughout the world [3]. Isolates of the 6th pandemic were almost exclusively of the O1 classical biotype, whereas the current (7th) pandemic is dominated by *V. cholerae* O1 El Tor biotype as the causative agent, a transition occurring between 1905 and 1961. The six pandemics previous to the current pandemic are considered to have originated in the Indian subcontinent, whereas the 7th pandemic strain was first isolated in the Indonesian island of

Sulawesi in 1961, and subsequently in Asia, Africa, and Latin America.

Over the last twenty years, several new epidemic lineages of *V. cholerae* O1 El Tor have emerged or re-emerged. In 1992, a new serogroup of *V. cholerae*, O139, was identified as the cause of epidemic cholera in India and Bangladesh [4]. That is, both *V. cholerae* O1 El Tor and O139 consistently have been isolated where the major cholera epidemics since 1992 have occurred, although the geographic location of *V. cholerae* O139 appears still to be restricted to Asia. Additionally, *V. cholerae* "hybrid" O1 El Tor variants that carry the classical type CTX prophage, or produce classical type cholera toxin subunit B have been repeatedly isolated in Bangladesh [5,6] and Mozambique [7]. These new variants have subsequently replaced the prototype 7th pandemic *V. cholerae* O1 El Tor strains in Asia and Africa, with respect to frequency of isolation from clinical cases of cholera.

It is clear that the dynamics of *V. cholerae*, like other enteric pathogens, can be attributed to extensive lateral gene transfer *via* transduction, conjugation, and transformation [2,8,9]. However, the evolutionary history of this bacterium remains to be documented. Here we compare the genome sequences of 23 *V. cholerae* strains, representing diverse serogroups that have been isolated at various times over the past 98 years from a variety of sources and geographical locations. We conclude that the current pandemic is caused by strains belonging to a single phyletic line, diversified mainly by lateral gene transfer occurring in the natural environment.

Phylogeny and gene content of *Vibrio cholerae.*

Phylogenetic analysis, accomplished using *ca.* 1.4 million bp of orthologous protein-coding regions for 23 *V. cholerae* strains (Table 1), revealed 12 distinct phyletic lineages. Strains belonging to non-O1/non-O139 serogroups from various sources showed substantial genomic diversity (Fig. 1a). In contrast, all *V. cholerae* serogroup O1 strains, except for two, comprised a monophyletic clade, designated *V. cholerae* core genome (CG) clade. Strains of both the 6th and 7th pandemics are concluded to have evolved from a common ancestor of this CG clade.

Twelve strains of the CG clade were further divided into two subgroups, as shown in the phylogenetic tree constructed using *ca.* 2.6 million bp alignment (Fig. 1a,b). The CG-1 subclade is comprised of most of the *V. cholerae* O1 El Tor strains and one *V. cholerae* O139 strain, whereas the CG-2 subclade contains strains of *V. cholerae* O1 classical and O37 serogroups. Interestingly, all clinical isolates associated with the current 7th cholera pandemic formed a very tight, monophyletic clade within the CG-1 subclade, which we have designated the 7th pandemic (7P) clade (Fig. 1b). *V. cholerae* O1 El Tor and O139 strains isolated from the Indian subcontinent and Africa epidemics during 1975 to 2004 are located in the 7P clade.

V. cholerae O1 El Tor N16961, the first *V. cholerae* strain to have been sequenced [10],

contains 3,908 ORFs, according to the RAST annotation server [11]. Based on a comprehensive ortholog detection using reciprocal comparison, we identified 2,432 "core" ORFs (66.9%) in all 23 *V. cholerae* strains, results similar to those reported for *Escherichia coli* (2,344 [12]and 2,865 [13]). Core gene ratios of the large and small chromosomes were 69.7 and 59.9% of the total genome of strain N16961, respectively. The higher core gene content obtained for the large chromosome implies a functional importance for genes in this chromosome. Heidelberg et al. [10] had shown that the large chromosome contained fewer genes coding for hypothetical proteins than the small chromosome.

Comparative genomic analysis was used to identify a set of non-redundant orthologues, yielding 6,953 gene sets of the pan-genome of *V. cholerae* based on 23 strains. The number of non-redundant genes found in 7P, CG-1 and CG clades were 4,327, 4,664 and 4,818, respectively. Overall, more genes were newly found when two strains showed further evolutionary distance (Supplementary Fig. 1b). On average, an increase of 206 new genes occurred each time a strain was added to the *V. cholerae* gene pool from which the pan-genome was calculated, if the strains that were added represented unique phyletic lines outside the CG group (Supplementary Fig. 1c). The number of additional genes per strain added is significantly higher than reported for *Streptococcus agalactiae* (27 genes) [14], but less than for *E. coli* (300 genes)[12].

O serogrouping in the context of genome evolution.

The lipopolysaccharide (LPS) of *V. cholerae* consists of three major regions: lipid A, core oligosaccharide (OS), and O antigen. *V. cholerae* synthesizes core OS and O antigen using *wav* and *wb** gene clusters, respectively [15,16]. Molecular phylogeny and genetic organization of the *wav* and *wb** gene clusters, located adjacently in the large chromosome, are summarized in Fig. 2 . Nesper *et al.* [15] surveyed *wav* gene clusters of 38 *V. cholerae* strains and categorized them into five types and, in this study, core OS types 1, 3, 4 and 5 were detected. Interestingly, *V. cholerae* biovar albensis VL426 contains a variant of type 3, which we have designated type 3b.

In contrast to the limited diversity observed in the *wav* gene cluster (5 major types), 11 different types of *wb** gene clusters were observed among the 23 strains. Phylogeny and genetic organization, based on the whole genome (Fig. 1), core OS, and O antigen gene clusters (Fig. 2a), clearly indicate both core OS and O antigen gene clusters are mobile *via* lateral gene transfer. The relatively stable gene order (synteny) of the core OS gene cluster indicates that it transfers as an entity. In contrast, the region coding for the O antigen is comprised of combinations of several smaller gene sets of different origin, leading to a remarkable diversity of the various O antigens seen in nature (Fig. 2b). This finding is in good agreement with a previous study [17] showing that the gene cluster coding for the O139 antigen is similar in *V. cholerae* serogroup O22, where substitution of a part of the cluster

occurred, but not a deletion.

Genome phylogeny (Fig. 1a) revealed that strains of O1 serogroup are found in three different phyletic lineages, namely the CG clade, and the *V. cholerae* O1 El Tor 12129(1) and TM11079-80 strains, in which the coding region for the O1 antigen is nearly identical. It is concluded that the O1 antigen phenotype arose at least three times in the evolution of *V. cholerae*. Furthermore, we hypothesize that the ancestor of the CG clade possessed a combination of the type 1 core OS and the O1 antigen gene clusters, giving rise to the present 12 *V. cholerae* CG strains, including the two *V. cholerae* non-O1 strains (V52 and MO10). The latter two became different serogroups by gene replacement, *via* lateral gene transfer, with strain V52 receiving both type 1 core OS and O37 antigen gene clusters from a *V. cholerae* O37 strain and *V. cholerae* MO10 receiving only the *V. cholerae* O139 antigen gene cluster from an unknown source, possibly a variant of the *V. cholerae* O22 serogroup [17].

The *V. cholerae* O1 strains not belonging to the CG group, *V. cholerae* 12129(1) and TM11079-80, are environmental isolates from Australia isolated in 1985 [18], and from Brazil, isolated in 1980 [19]. They showed the typical El Tor phenotype, but unlike other *V. cholerae* O1 El Tor strains in the CG-1 subclade, lack the two major virulence-related genomic islands, i.e., CTX prophage containing cholera toxin genes (*ctx*AB) and *Vibrio* pathogenicity island-1 (VPI-1) containing genes for biosynthesis of the toxin co-regulated pilin (TCP). By comparing genome phylogenies based on the whole genome (Fig. 1) and gene clusters coding

for the core OS (Fig. 2a) and O1 antigen (Fig. 2c), it is clear that genesis of these non-toxigenic *V. cholerae* O1 El Tor strains can be attributed to independent lateral gene transfer events, perhaps transfer of only the O1 antigen gene cluster, but not the core OS region. The O1 antigen gene cluster itself appears to be highly conserved, since only 66 nucleotides were found to vary within the 19,503 bp multiple sequence alignment spanning 19 orthologues among the 11 *V. cholerae* O1 strains sequenced in this study. In any case, the phylogenetic analysis suggests, that donors of the O1 antigen gene cluster of *V. cholerae* 12121(1) and TM11079-80 were most likely members of the CG-2 subclade encompassing *V. cholerae* O1 classical biotype (Fig. 2c).

Thus, at least four O serogroup conversions, were detected, i.e., conversion from non-O1 to O1 (twice), O1 to O139, and O1 to O37, among the 23 *V. cholerae*. Earlier, several non-genomics based studies suggested such conversions take place in nature [16,19,20], and chitin-induced natural transformation has been proposed as a mechanism in the natural environment [21]. Subsequently, *V. cholerae* O1 to O139 serogroup conversion by a single-step exchange of large fragments of DNA was demonstrated in a microcosm experiment [9], clearly supported by the conclusion of this study that O serogroup conversion occurs frequently in nature. Mobility of the O phenotype in *V. cholerae* was first proposed by Colwell et al. [22], and the cumulative results of both *in vivo* and *in vitro* experiments, as well as the genomic evidence, are compelling. Given the inconsistency between O serogroup typing and genome-

based phylogeny, it is concluded that, at the very least, the term "O1 El Tor" is both misleading and inaccurate for describing a set of phylogenetically coherent *V. cholerae* strains, especially in light of the frequency of serogroup conversion. Therefore, we propose a new terminology based on genome sequence; namely the core genome (CG) clade, CG-1 and CG-2 subclades, and 7th pandemic (7P) clade, to describe homologous intraspecific groups of *V. cholerae* (Fig. 1).

Virulence-associated prophage and genomic islands within the context of the genome.

V. cholerae possesses several known virulence factors, of which the cholera toxin (CT) and toxin-coregulated pilus (TCP) are considered to be the most significant. Genes coding for CT (*ctx*AB) are part of a temperate filamentous bacteriophage CTX□ [8] that can be incorporated into both chromosomes of *V. cholerae* at specific positions. The CTX□ genes were found to be present in members of the CG clade, except for *V. cholerae* NCTC 8457 and 2740-80. Among strains not belonging to the CG clade, only *V. cholerae* serogroup O141 (V51) contains this prophage.

The CTX□ found in classical and El Tor biotypes differs in the sequence of their repressor gene, *rst*R, and are classified as CTX□^{Class} and CTX□$^{El\ Tor}$, according to the biotype of the host [23]. From the genome sequences, it was found that CTX□^{Class} is not restricted to the classical biotype, but is also widely distributed in *V. cholerae* O1 El Tor and O141 strains

(Fig. 3). Given that *V. cholerae* O1 El Tor MAK 757, a clinical strain isolated in 1937, has this type of prophage, correlation of host biotype and prophage type is not considered significant. A recent study[24] showing the infection of CTX□^{Class} to *V. cholerae* non-O1 supports our finding.

Chromosomal attachment sites for CTX□ are known to harbor other genetic elements, including toxin-linked cryptic (TLC), RS1 elements, and VSK(=pre-CTX) prophages [25,26]. Additionally, we have discovered five genomic islands (GI-19, GI-22, GI-33, GI-43, GI-48) in the region of the CTX□ attachment sites on both chromosomes. In total, nine distinct genetic elements were found in these regions, where they appear in different combinations (Fig. 3). Seven strains possess GI-19 in either chromosome, which is similar but not identical to KSF-1□□, discovered in an environmental *V. cholerae* strain [27]. It is evident that more bacteriophages/genetic elements are located in the CTX□ attachment regions of CG strains than non-CG strains. The ability to harbor more, especially the toxigenic, bacteriophage-like elements, in these regions in CG strains may explain why only members of the CG strains have been agents of cholera pandemics over the last century, namely during the 6th and 7th pandemics. We found no two toxigenic (CTX □-harboring) strains with an identical GI organization and combination, except for two "hybrid" strains (the only 7P members harboring CTX□□^{Class}). It is evident from Fig. 3 that the two CTX□□ attachment sites serve as an engine of genetic diversity for *V. cholerae* CG clade.

Genes coding for the toxin-coregulated pilus (TCP) are part of a genomic island, VPI-1, that is present in all CG strains. Among the non-CG strains, only *V. cholerae* O141 V51 contained VPI-1 but with less sequence similarity. Since TCP serves as a receptor for CTX□, it explains why only this strain, among all of the non-CG strains, possesses CTX□. Results of phylogenetic analysis based on the 24 genes of VPI-1 suggest that the original GI of *V. cholerae* NCTC 8457 has subsequently been replaced by a VPI-1 of a non-CG strain (Supplementary Fig. 2). Interestingly, GI-47, but not VPI-1, was found in the same genomic region in strains MZO-3, 1587, MZO-2, and VL426. This cassette-like property of GI mobility was also observed for the other known pathogenicity islands, including VPI-2, VSP-1, and VSP-2 (Supplementary Table 3).

Extensive lateral gene transfer in *V. cholerae*.

Since it is generally accepted that lateral gene transfer plays an important role in the evolution of pathogenic bacteria, *V. cholerae* can serve as a useful paradigm. In this study, we define a GI as a genomic region containing five or more ORFs, where transfer, but not deletion, is obvious from comparison of the genome phylogeny and its presence/absence among test strains. A total of 73 GIs were identified (Supplementary Table 2) and their chromosomal locations are shown in Fig. 4. As discussed earlier, with respect to GIs associated with O antigen biosynthesis, CTX□, VPI-1,2 and VSP-2, a total of 13 (eight in the

large and five in the small chromosome) genomic regions were found to have a cassette-like property, whereby different GIs occupy the same or similar region (Supplementary Table 3). Most GIs were found to be singletons in a given genome, although two (GI-12, GI-21) were present as four and two copies, respectively. Thus, we conclude that genetic diversity of the species *V. cholerae* derives predominantly by lateral gene transfer, of which several such transfers appear to be cassettes.

Genomic definition of *V. cholerae* core genome (CG) clade and pandemic strains.

The *V. cholerae* CG clade, with both 6th and 7th pandemic strains, is defined by gene content as follows. Twenty-seven genes were present exclusively in the genomes of the CG strains, but only five genes are unique to the CG-1 subclade. Four of these (VCA0198-VCA0201) comprise a genomic island (GI-5) on the small chromosome, including genes coding for cytosine-specific DNA methyltransferase [28] and hypothetical proteins, adjacently located to the IS1004 transposase gene. The 7P strains are differentiated in harboring two unique GIs, the *Vibrio* seventh pandemic island-1 (VSP-1) and VSP-2, first discovered by microarray analysis [29]. In addition to the 7P strains, a variant of VSP-1 was found in *V. cholerae* biovar albensis VL426 (Supplementary Fig. 3). Similarly, VSP-2 like GIs were detected in three non-CG strains (TMA21, O39 MZO-3, O135 RC385). Interestingly, a similar GI was also detected in *Vibrio vulnificus* YJ016 and *Vibrio splendidus* 12B01, suggesting that VSP-2 may

be widespread in its distribution among vibrios (Supplementary Fig. 4). It should be noted that stability of these well known pathogenicity islands among 7P members is questionable since most of VPI-2 and VSP-2 were found to be deleted in MO10 and CIRS 101, respectively.

V. cholerae contains a super-integron, a large integron island (gene capture system), in the small chromosome (~120 Kbp), a region comprising predominantly hypothetical genes and proposed as a source of genetic variation [10]. All *V. cholerae* strains examined in this study have this integron, a source of much of the variation in gene content (Supplementary Fig. 5). Interestingly, if this region is excluded, all six members of the 7P clade have an identical gene content, with exception of a few genomic islands, including those found in the CTX attachment region. An SXT element, belonging to a family of conjugative transposon-like mobile genetic elements, encodes multiple antibiotic resistance genes and is present only in *V. cholerae* MO10, CIRS 101, MJ-1236, and B33, but not in the other *V. cholerae* strains. *V. cholerae* O139 MO10 differs from other members of the 7P clade in having an O139 antigen specific genomic island, a finding that strongly supports the conclusion of several previous studies, namely that *V. cholerae* O139 derives from a 7th pandemic *V. cholerae* O1 El Tor strain [30]. No other *V. cholerae* O139-specific genes were found in *V. cholerae* MO10.

The "hybrid" strains, possessing an El Tor biotype phenotype, but classical biotype CTX, were isolated during current cholera epidemics in Asia and Africa [6,7]. Two hybrid

strains (B33 and MJ-1236) share a virtually identical genome backbone. Among 3,587,239 bp of orthologous protein-coding region, only 106 nucleotide positions are different and the only significant difference is the presence of a *V. cholerae* MJ-1236 specific 19,729 bp genomic island (GI-12). This GI occurs four times as an almost identical sequence in the large chromosome, with 14 genes including those coding for the putative phage integrase and type I restriction-modification system, implying that it is probably a recently introduced novel temperate bacteriophage. It is not clear why these hybrid strains outcompete other *V. cholerae* O1 El Tor/O139 in the clinical setting, but a key to the puzzle surely lies in the differences among closely related strains, i.e., tandem copies of CTXΦ^{Class}, GI-14 and single nucleotide polymorphisms. In addition to these hybrid clones, *V. cholerae* O1 El Tor strains producing a more virulent type of cholera toxin subunit B, i.e., classical, repeatedly have been isolated from patients in Asia and Africa [6]. The genome sequence of a representative of this newly emerged group, i.e. *V. cholerae* CIRS 101, reveals that these strains also have a typical 7P gene content, but with CTX$\Phi^{El\ Tor}$, not CTXΦ^{Class}, albeit expressing the classical type subunit B protein (Fig. 3).

The comparative genomics of phylogenetically diverse strains has permitted analysis of the mechanism by which current 7th pandemic clones may have arisen. An highly conserved gene content, synteny, and significant similarity among the six strains of the 7P clade indicate that these *V. cholerae* strains share an almost identical genome "backbone",

having evolved very recently from a common ancestral strain. An hypothetical evolutionary pathway proposed for *V. cholerae* (Fig. 5), with GI migration matched to a genome-based phylogenetic tree, allows the conclusion that the ancestor for the 7P clade was a *V. cholerae* O1 El Tor strain containing several GIs (VPI-1,2, GI-1 to GI-10), receiving VSP-1, VSP-2 and GI-11(=Kappa prophage) by lateral gene transfer, and finally giving rise to the contemporary *V. cholerae* O1 El Tor and O139 strains. It is interesting to note that such an hypothetical ancestral strain shows a gene content similar to *V. cholerae* O1 El Tor BX330286 isolated from a water sample collected in Australia in 1986, a geographic location near Indonesia where the first 7th pandemic *V. cholerae* O1 El Tor was reported in 1961.

Mechanism of *Vibrio cholerae* evolution.

Of the few human pathogens for which multiple genomes have been fully sequenced [14,12,31], *V. cholerae* provides a unique opportunity to elucidate evolutionary mechanism(s) that can be generally applied to other bacterial species, since *V.cholerae* is both highly pathogenic for humans and a successful inhabitant of the natural environment worldwide. It is indigenous to estuarine environments of both cholera epidemic and non-epidemic countries [3].

Unlike *Samonella* Typhi and *Bacillus anthracis*, which are special cases of bacterial species showing clonal properties, *V. cholerae,* with *Streptococcus agalactiae* and *Escherichia coli*, offers a prime example of the important role of lateral gene transfer in the

evolution of a bacterial species. The transition from 6th to 7th cholera pandemic genome type is concluded to result from a change of causal agents of *V. cholerae* O1 classical to *V. cholerae* O1 El Tor biotype. We propose the term "shift" for the event occurring between two distinct phyletic lineages (Fig. 1a). In contrast, the present cholera global pandemic is ascribed to a clonal shift/change among 7P strains, e.g. emergence of *V. cholerae* O139, *V. cholerae* O1 El Tor hybrid and *V. cholerae* O1 El Tor with an altered cholera toxin subunit B. These represent transitions among genetically nearly identical clones, with a few different GIs, for which we propose the term "drift". Much as in the case of the influenza viruses, cholera bacteria undergo a shift/drift cycle over time, though the drift in *V. cholerae* is derived mainly from lateral gene transfer, but not mutation, most likely occurring in the natural environment, where the reservoir of *V. cholerae* occurs in association with its plankton hosts [3,32].

The present cholera global pandemic is concluded to have been initiated by multiple descendants of a *V. cholerae* O1 El Tor ancestor, diversified and continuously rapidly evolving, mainly *via* lateral gene transfer and most likely driven by environmental factors. Most importantly, the common genome backbone and variable genomic islands of the 7P clade of *V. cholerae* require that a reevaluation be done of the epidemiological practice that employs serogroups as the primary marker for *V. cholerae.* The so-called pandemic clones, instead, should be defined by gene content, the description of which offers significantly

greater potential for the development of reliable and useful diagnostics, vaccines, and therapeutics for cholera. Without doubt, more variants of the 7P clade, as a result of "drift", will be encountered in the future, yielding new serogroups (other than O1 and O139) and phenotypic combinations. Public health workers will be unprepared if the evolution of this species remains unappreciated as an ongoing process in the natural environment, where *V. cholerae* is autochthonous and plays an important role in the nutrient cycles of the natural aquatic ecosystem.

METHODS SUMMARY

Nearly complete genome sequences were obtained from a blend of Sanger, 454 and Solexa sequences using standard protocols used in the DOE Joint Genome Institute. Gene-finding and annotation were achieved using the RAST server [11]. Pairwise genome by genome comparison was carried out using BLASTN, BLASTP and TBLASTX analyses of each ORFs. Additionally, orthologous regions were identified by pairwise global alignment, and used for translation-frame-independent means of comparison. Orthologs and paralogs were differentiated by reciprocal comparison. Sets of orthologous genes were aligned separately by CLUSTALW2 and subsequently concatemerized to generate phylogenetic trees using Kimura-2-parameter model [33] and neighbor-joining method [34]. Identification of genomic islands were carried out manually by comparing ortholog content among test strains and phylogenetic tree based on genome sequence.

Acknowledgements

This study was supported by the KOSEF National Research Laboratory Program (Grant No. R0A-2005-000-10110-0 to J.C.), National Institutes of Health (Grant No. 1RO1A139129-01 to R.R.C.), National Oceanic and Atmospheric Administration, Oceans and Human Health Initiative (Grant No. S0660009 to R.R.C.), IC Post-Doctoral Fellowship Program (to C.J.G) and the Korean and Swedish governments (to IVI). Funding for sequencing was provided by the Office of the Chief Scientist (USA).

METHODS

Genome sequencing. Draft sequences were obtained from a blend of Sanger and 454 sequences and involved paired end Sanger sequencing on 8kb plasmid libraries to 5X coverage, 20X coverage of 454 data, and optional paired end Sanger sequencing on 35kb fosmid libraries to 1-2X coverage (depending on repeat complexity). To finish the genomes, a collection of custom software and targeted reaction types were used. In addition to targeted sequencing strategies, Solexa data in an untargeted strategy were used to improve low quality regions and to assist gap closure. Repeat resolution was performed using in-house custom software (dupFinisher, C. Han, unpublished). Targeted finishing reactions included transposon bombs [35], primer walks on clones, primer walks on PCR products, and adapter PCR reactions. Gene-finding and annotation were achieved using the RAST server [11].

Comparative genomics. Genome to genome comparison was performed using three approaches, since completeness and quality of nucleotide sequences varied from strain to strain in the set examined in this study. Firstly, nucleotide sequences as whole contigs were directly aligned using the mummer program [36]. Secondly, ORFs of a given pair of genomes were reciprocally compared each other, using the BLASTN, BLASTP and TBLASTX programs (ORF-dependent comparison). Thirdly, a bioinformatic pipeline was developed to identify homologous regions of a given query ORF. Initially, a segment on target contig homologous to a query ORF was identified using the BLASTN program. This potentially

homologous region was then expanded in both directions by 2,000 bp. Then, nucleotide sequences of the query ORF and selected target homologous region were aligned using a pairwise global alignment algorithm [37], and the resultant matched region in subject contig was extracted and saved as a homolog (ORF-independent comparison). Orthologs and paralogs were differentiated by reciprocal comparison. In most cases, both ORF-dependent and -independent comparisons yielded the same orthologs, though ORF-independent method performed better for draft sequences of low quality, in which sequencing errors, albeit rare, hampered identification of correct ORFs.

Identification and annotation of genomic islands. In this study, we defined genomic islands (GIs) as a continuous array of five or more ORFs that were found to be discontinuously distributed among genomes of test strains. Correct transfer or insertion of GIs was readily differentiated from deletion event by comparing genome-based phylogenetic tree and full matrices showing pairwise detection of orthologous genes between test strains. Identified GIs were designated, and annotated using the BLASTP search of its member ORFs against Genbank NR database.

Phylogenetic analyses based on genome sequences. A set of orthologues for each ORF of *V. cholerae* N16961 was obtained for different sets of strains, and then aligned individually using the CLUSTALW2 [38] program. The resultant multiple alignments were concatenated to generate genome scale alignments, which were subsequently used to reconstruct the

neighbor-joining phylogenetic tree [34]. The evolutionary model of Kimura [33] was used to generate the distance matrix. The MEGA program [39] was used for phylogenetic analysis.

FIGURE LEGENDS

Figure 1. Neighbor-joining trees showing phylogenetic relationships of 23 *V. cholerae* strains representing diverse serogroups. (a) All *V. cholerae* strains based on 1,676 ORFs (1,370,469 bp) (b) core genome (CG) clade based 2,663 ORFs (2,567,393 bp) (c) 7^{th} pandemic (7P) clade based on 3,364 ORFs (3,291,577 bp). Bootstrap supports, as percentage, are indicated at the branching points. Bars represent the numbers of substitution per site, respectively. Only orthologous genes showing >95% nucleotide sequence similarity to those of *V. cholerae* N16961 were selected. The tree was rooted using *Vibrio vulnificus* YJ016 and *Vibrio parahaemolyticus* RIMD 2210633.

Figure 2. (a) Neighbor-joining tree of gene cluster for core oligosaccharide (OS) biosynthesis based on 3,982 bp long alignment, which was generated from four ORFs (VC0227, VC0234, VC0236, VC0239) commonly present in all 23 *V. cholerae* strains. (b) Organization of *wav* and *wb** gene clusters of *V. cholerae*. The homologous regions in the O antigen coding gene cluster are indicated in the same colors. The nomenclature of *wav* gene cluster coding for core oligosaccharide is according to Nesper et al. Strain VL426 has a variant of type 3 which we designate type 3b in this study. (c) Neighbor-joining tree based on 19 genes (VC0241-VC0254, VC0259-VC0263) coding for O1 antigen. The numbers at nodes indicate the bootstrap supports

and bars indicate the number of substitutions per a nucleotide position.

Figure 3. Schematic representation of various prophages and genetic elements present in the target regions of CTX□ insetion. The numbers in parentheses are year of isolation and source (C, clinical; E, Environment). * TLC, El Tor type CTX □, RS1 element are found, but no positional information can be obtained from assemblies. † Classical type CTX □ and RS1 are present, but no positional information can be obtained.

Figure 4 . Genomic representation of genomic islands of both *V. cholerae* chromosomes. The two circles in the middle represent the genes in *V. cholerae* O1 El Tor N16961. The inner circle indicates genomic islands found in strain N16961, whereas the outer circles are those absent in strain N16961.

Figure 5. Proposed hypothetical evolutionary pathway of the *V. cholerae* species. Probable insertions and deletions of genomic islands (Supplementary Table 2) found in 23 *V. cholerae* strains are indicated by black and red arrows, respectively, along the phylogenetic tree based on genome sequence data. Hypothetical ancestral strains are indicated by open circles.

REFERENCES

1 Chatterjee, S. N. & Chaudhuri, K. Lipopolysaccharides of *Vibrio cholerae.* I. Physical and chemical characterization. *Biochim Biophys Acta* **1639**, 65-79 (2003).

2 Kaper, J. B., Morris, J. G., Jr. & Levine, M. M. Cholera. *Clin Microbiol Rev* **8**, 48-86 (1995).

3 Colwell, R. R. Global climate and infectious disease: the cholera paradigm. *Science* **274**, 2025-2031 (1996).

4 Ramamurthy, T. *et al.* Emergence of novel strain of Vibrio cholerae with epidemic potential in southern and eastern India. *Lancet* **341**, 703-704 (1993).

5 Nair, G. B. *et al.* New variants of Vibrio cholerae O1 biotype El Tor with attributes of the classical biotype from hospitalized patients with acute diarrhea in Bangladesh. *J Clin Microbiol* **40**, 3296-3299 (2002).

6 Nair, G. B. *et al.* Cholera due to altered El Tor strains of Vibrio cholerae O1 in Bangladesh. *J Clin Microbiol* **44**, 4211-4213, doi:JCM.01304-06 [pii] 10.1128/JCM.01304-06 (2006).

7 Ansaruzzaman, M. *et al.* Cholera in Mozambique, variant of Vibrio

cholerae. *Emerg Infect Dis* **10**, 2057-2059 (2004).

8 Waldor, M. K. & Mekalanos, J. J. Lysogenic conversion by a filamentous phage encoding cholera toxin. *Science* **272**, 1910-1914 (1996).

9 Blokesch, M. & Schoolnik, G. K. Serogroup conversion of Vibrio cholerae in aquatic reservoirs. *PLoS Pathog* **3**, e81 (2007).

10 Heidelberg, J. F. *et al.* DNA sequence of both chromosomes of the cholera pathogen Vibrio cholerae. *Nature* **406**, 477-483 (2000).

11 Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**, 75 (2008).

12 Rasko, D. A. *et al.* The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. *J Bacteriol* **190**, 6881-6893, doi:JB.00619-08 [pii]10.1128/JB.00619-08 (2008).

13 Chen, S. L. *et al.* Identification of genes subject to positive selection in uropathogenic strains of Escherichia coli: a comparative genomics approach. *Proc Natl Acad Sci U S A* **103**, 5977-5982, doi:0600938103 [pii]10.1073/pnas.0600938103 (2006).

14 Tettelin, H. *et al.* Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome".

Proc Natl Acad Sci U S A **102**, 13950-13955, doi:0506758102 [pii]10.1073/pnas.0506758102 (2005).

15 Nesper, J. *et al.* Comparative and genetic analyses of the putative Vibrio cholerae lipopolysaccharide core oligosaccharide biosynthesis (wav) gene cluster. *Infect Immun* **70**, 2419-2433 (2002).

16 Li, M., Shimada, T., Morris, J. G., Jr., Sulakvelidze, A. & Sozhamannan, S. Evidence for the emergence of non-O1 and non-O139 Vibrio cholerae strains with pathogenic potential by exchange of O-antigen biosynthesis regions. *Infect Immun* **70**, 2441-2453 (2002).

17 Yamasaki, S. *et al.* The genes responsible for O-antigen synthesis of vibrio cholerae O139 are closely related to those of vibrio cholerae O22. *Gene* **237**, 321-332, doi:S0378111999003443 [pii] (1999).

18 Safa, A. *et al.* Multilocus genetic analysis reveals that the Austrailian strains of *Vibrio cholerae* O1 are similar to the pre-seventh pandemic strains of the El Tor biotype. *J. Med. Microbiol.* **58**, 105-111 (2009).

19 Farfan, M., Minana, D., Fuste, M. C. & Loren, J. G. Genetic relationships between clinical and environmental *Vibrio cholerae* isolates based on multilocus enzyme electrophoresis. *Microbiology* **146 (Pt 10)**, 2613-2626 (2000).

[20] Bik, E. M., Gouw, R. D. & Mooi, F. R. DNA fingerprinting of Vibrio cholerae strains with a novel insertion sequence element: a tool to identify epidemic strains. *J Clin Microbiol* **34**, 1453-1461 (1996).

[21] Meibom, K. L., Blokesch, M., Dolganov, N. A., Wu, C. Y. & Schoolnik, G. K. Chitin induces natural competence in Vibrio cholerae. *Science* **310**, 1824-1827 (2005).

[22] Colwell, R. R., Huq, A., Chowdhury, M. A., Brayton, P. R. & Xu, B. Serogroup conversion of *Vibrio cholerae*. *Can J Microbiol* **41**, 946-950 (1995).

[23] Davis, B. M., Moyer, K. E., Boyd, E. F. & Waldor, M. K. CTX prophages in classical biotype Vibrio cholerae: functional phage genes but dysfunctional phage genomes. *J Bacteriol* **182**, 6992-6998 (2000).

[24] Udden, S. M. *et al.* Acquisition of classical CTX prophage from Vibrio cholerae O141 by El Tor strains aided by lytic phages and chitin-induced competence. *Proc Natl Acad Sci U S A* **105**, 11951-11956, doi:0805560105 [pii] 10.1073/pnas.0805560105 (2008).

[25] Rubin, E. J., Lin, W., Mekalanos, J. J. & Waldor, M. K. Replication and integration of a Vibrio cholerae cryptic plasmid linked to the CTX

prophage. *Mol Microbiol* **28**, 1247-1254 (1998).

26 Faruque, S. M. *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* **104**, 5151-5156 (2007).

27 Faruque, S. M. *et al.* CTXphi-independent production of the RS1 satellite phage by Vibrio cholerae. *Proc Natl Acad Sci U S A* **100**, 1280-1285, doi:10.1073/pnas.0237385100 [pii] (2003).

28 Banerjee, S. & Chowdhury, R. An orphan DNA (cytosine-5-)-methyltransferase in Vibrio cholerae. *Microbiology* **152**, 1055-1062, doi:152/4/1055 [pii]
10.1099/mic.0.28624-0 (2006).

29 Dziejman, M. *et al.* Genomic characterization of non-O1, non-O139 *Vibrio cholerae* reveals genes for a type III secretion system. *Proc Natl Acad Sci U S A* **102**, 3465-3470 (2005).

30 Karaolis, D. K., Lan, R. & Reeves, P. R. Molecular evolution of the seventh-pandemic clone of Vibrio cholerae and its relationship to other pandemic and epidemic V. cholerae isolates. *J Bacteriol* **176**, 6199-6206 (1994).

31 Holt, K. E. *et al.* High-throughput sequencing provides insights into

genome variation and evolution in Salmonella Typhi. *Nat Genet* **40**, 987-993, doi:ng.195 [pii]10.1038/ng.195 (2008).

[32] Constantin de Magny, G. *et al.* Environmental signatures associated with cholera epidemics. *Proc Natl Acad Sci U S A* **105**, 19676-17681 (2008).

[33] Kimura, M. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *J Mol Evol* **16**, 111-120 (1980).

[34] Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol Biol Evol* **4**, 406-425 (1987).

[35] Goryshin, I. Y. & Reznikoff, W. S. Tn5 in vitro transposition. *J Biol Chem* **273**, 7367-7374 (1998).

[36] Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12 (2004).

[37] Myers, E. W. & Miller, W. Optimal alignments in linear space. *Comput Appl Biosci* **4**, 11-17 (1988).

[38] Larkin, M. A. *et al.* Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).

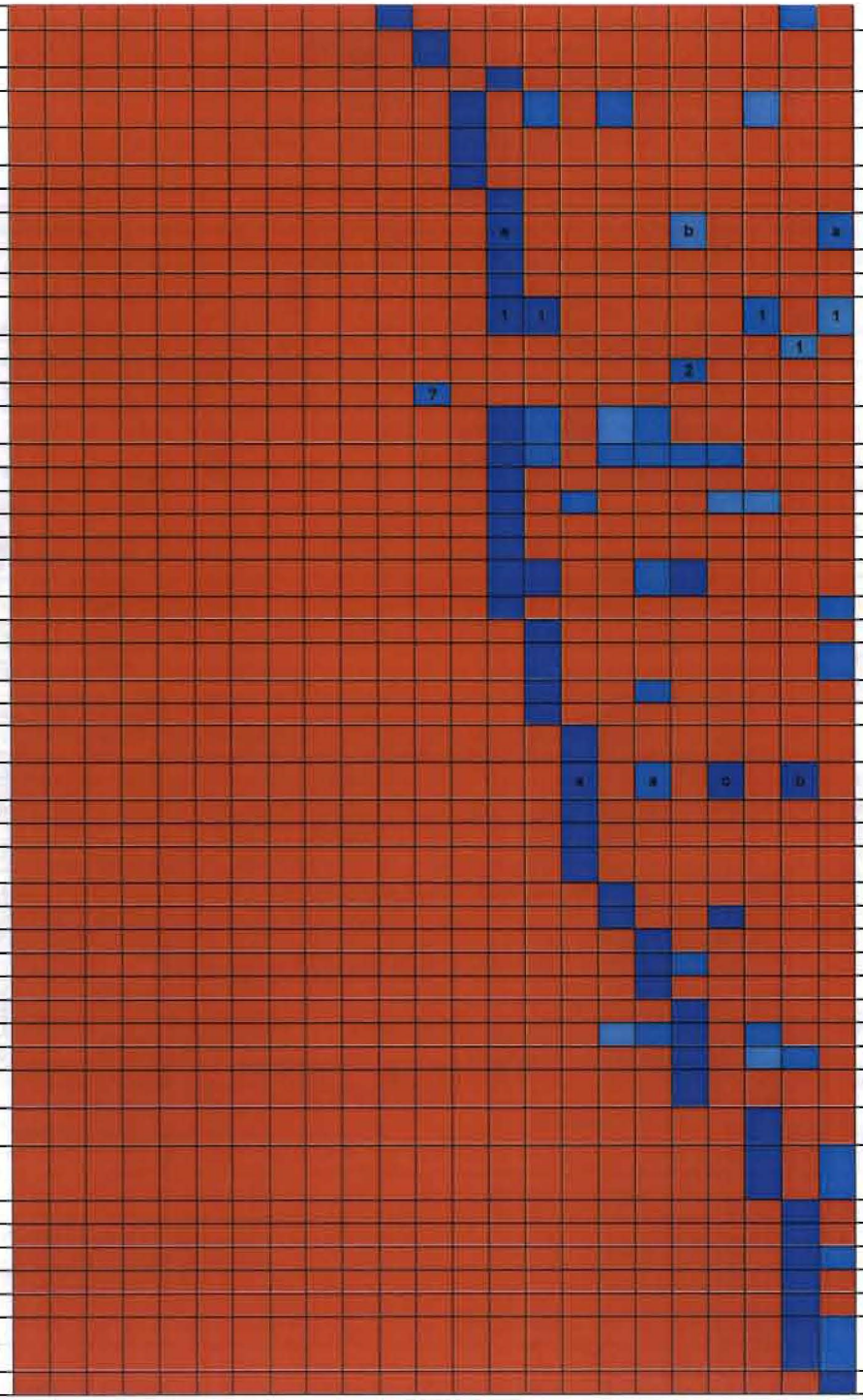
[39] Kumar, S., Nei, M., Dudley, J. & Tamura, K. MEGA: a biologist-centric software for evolutionary analysis of DNA and protein sequences. *Brief*

Bioinform **9**, 299-306 (2008).

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

Supplementary Table 2. Genomic islands identified among 23 *Vibrio cholerae* strains in this study.

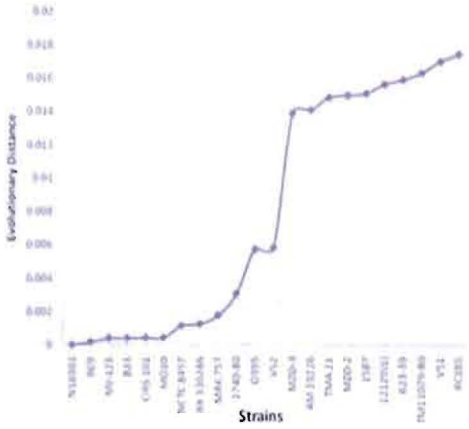
GI- designation	Chromosome	ORFs*	Number of ORFs	Insertion region based on N16961 locus	VCN16961	VCRC9	VCMJ-236	VCB33	VCCIRS101	VCMO10	VCBX330286	VC8457	VC274080	VCMAK757	VCO395	VCV52	VC12129	VCTM11079	VCTMA21	VCMZO3	VCAM19226	VC1587	VC62339	VCMZO2	VCV51	VCVL426	VCRC385	Description/Functional Annotation
VSP-1	1	VCN16961.173~83	7-12	VC0174-VC0186																								Vibrio seventh pandemic island-1 (VSP-1)
VSP-2	1	VCN16961.485~508	21-30	VC0489-VC0517																								Vibrio seventh pandemic island-2 (VSP-2)
VPI-1	1	VCN16961.800~835	33-41	VC0809-VC0848																								Vibrio pathogenicity island-1 (VPI-1)
VPI-2	1	VCN16961.1735~1784	49-51	VC1757-VC1810																								Vibrio pathogenicity island-2 (VPI-2)
SXT	1	VCMJ1236.2645~2735	89-91	VC0659																								SXT element
CTX	1	VCN16961.1436~1448	9	VC1451-1465																								Cholera toxin prophage (CTX)
	2	VCMJ1236.3547~3556	10	VCA0569-VCA0570																								
TLC	1	VCN16961.1449~1460	13-14	VC1464-VC1479																								Cryptic plasmid linked to the CTX prophage
GI-1	1	VCN16961.1379~1388	11	VC1393-VC1405																								Motility and chemotaxis
GI-2	1	VCN16961.1564~1568	3-5	VC1582-VC1588																								Oxidative stress response
GI-3	1	VCN16961.1725~1729	4-5	VC1747-VC1753																								Membrane proteins
GI-4	1	VCN16961.1795~3954	8-9	VC1819-VC1828																								Carbohydrates (PTS system)
GI-5	2	VCN16961.2939~2943	4-5	VCA0197-VCA0204																								Site-specific DNA-methyltransferase
GI-6	2	VCN16961.3018~3023	5-6	VCA0280b-VCA0287																								Putative prophage
GI-7	2	VCN16961.3397~3403	4-7	VCA0666-VCA0674a																								Sodium-solute symporter/Sugar transporter
GI-8	2	VCN16961.3517~3522	5-6	VCA0789-VCA0796																								Transposable element
GI-9	2	VCN16961.3577~3584	7-8	VCA0849-VCA0858																								Autolysin sensor kinase/ABC-type transport system
GI-10	2	VCN16961.3605~3611	4-9, 19(b)	VCA0877-VCA0885																								Integrase/Non-hemolytic enterotoxin lytic component L1
GI-11	1	VCRC9.1391~1432	34-45	VC2187-VC2188																								Kappa prophage
	2	VC8457.1741~1785		VCA0197																								
GI-12	1	VCMJ1236.166~180	14	VC0425-VC0424																								HipA protein/ Type I restriction-modification system /Putative prophage
	1	VCMJ1236.1487~1500		VC1860-VC1859																								
	1	VCMJ1236.1594~1607		VC1761-VC1760																								
	1	VCMJ1236.2022~2035		VC1299																								
GI-13	2	VCMJ1236.3419~3427	9	VCA0697-VCA0695																								VSP-1 like element
GI-14	2	VCMJ1236.3632~3648	12-15	VCA0493-VCA0490																								Hypothetical proteins
	2	VCO395.3481~3492		VCA0480-VCA0481																								
GI-15	1	VCB33.3440~3443/3669~3681	18	VC0002-VC0003																								Phage/Conjugation/Endonuclease
GI-16	1	VCMO10.2722~2729	8	VC1451-VC1479?																								Vibrio VSK(=pre-CTX) prophage
	2	VCAM19226.3563-3566/2157~2159		VCA0569-VCA0570																								
GI-17	2	VC12129.3140~3180	39-46	VCA0885																								Putative prophage
GI-18	1	VC8457.401~409	9	VC0032																								Uropathogenic specific protein
GI-19	1	VC8457.942~955	9-14	VC1479-VC1451																								KSF-1□ -like prophage
	2	VC12129.2530~2540		VCA0569-VCA0570																								
GI-20	1	VC8457.1907~1948/3496-3498	45	VC2030																								Mu-like prophage
GI-21	1	VC274080.1654~1702	42-52	VC2634																								Mu-like prophage
	1	VCO395.631~680		VC1129-VC1130																								
	1	VCO395.719~768		VC1168																								
	2	VC12129.3336~3377		VCA0707																								
	2	VCTM11079.2779~2830		VCA1090-VCA1091																								
	?	VCTMA21.2153~2203/727~748		N.D.																								
	1	VCV51.3149~3198		VC0938																								
GI-22	1	VCMAK757.2724~2729	6	VC1463-VC1465																								Vibrio Vf33-like prophage
GI-23	1	VCO395.2280~2300	21	VC0156-VC0154																								Putative prophage

GI-24	1	VC0395.2577~2590	14	VC0289-VC0290		Putative prophage (CRISPR-associated proteins)
GI-25	1	VCV52.193~204/2707~2719/3459~3470	12-13	VC1349		Mu-like prophage
	1	VCTM11079.342~353		VC0048		
GI-26	1	VC12129.588~621	34-41	VC1773-VC1758		Type III secretion system/ Accessory colonization factor
GI-27	1	VC12129.1440~1457	18	VC0847-VC0807		Type II restriction-modification system/DNA repair protein RadC/Helicase
GI-28	1	VC12129.1922~1931	10	VC0373		Putative prophage
GI-29	1	VCTM11079.156~164	9	VC0398-VC0397		Putative prophage
GI-30	1	VCTM11079.287~297	11	VC0002-VC0003		Type I restriction-modification system/Virulence protein/Integrase
GI-31	1	VCTM11079.930~942	13	VC0487-VC0488		Transposable element
GI-32	1	VCTM11079.1205~1224	20	VC0209-VC0208		Putative prophage
GI-33	1	VCTM11079.643~644/1698~1704	9	VC1451-VC1479		Transposase/RstA1
	1	VCVL426.2545~2554		VC1451-VC1333		
	2	VC62339.506~510		VCA0569-VCA0570		
	?	VCV52.508~511		N.D.		
GI-34	1	VCTM11079.1925~1932	6-8	VC2714-VC2715		Beta-hexosaminidase (EC 3.2.1.52) /Ribose ABC transport system
GI-35	1	VCTMA21.2626~2632	7-9	VC1910-VC1911		ABC-type oligopeptide transport system
GI-36	1	VCTM11079.2103~2150	48	VC2041-VC2042		Putative prophage
GI-37	2	VCTM11079.2679~2686	2-8	VCA0990-VCA0994		Transposable element
GI-38	2	VCTM11079.3040~3062	23	VCA0197-VCA0204		Putative prophage
GI-39	2	VCTM11079.3090~3095	6	VCA0235-VCA0236		Transposable element
GI-40	2	VCTM11079.3279~3290	11-12	VCA0607-VCA0608		Maltose,maltodextrin ABC transporter/Cell wall synthesis
GI-41	1	VCTM11079.3535~3561	27	VC0847-VC0807		Putative prophage
GI-42	1	VCTMA21.24~33	10	VC0290-VC0289		Putative prophage
GI-43	2	VCTMA21.575~586	12	VCA0570-VCA0569		DNA polymerase III, alpha subunit (gram-positive type)/transcriptional regulator, XRE family
GI-44	2	VCTMA21.1759~1764	6	VCA1026-VCA1027		Carbohydrates metabolism
GI-45	1	VCTMA21.3543~3559	17	VC0379(-VC0378)		Putative prophage
GI-46	1	VCMZO3.1058~1126/2449~2473	94	VC1810-VC1757		VPI-1 like element
GI-47	1	VCMZO3.2003~2011/2091~2092	11-34	VC0809-VC0848		Cell division protein ftsK/Exonuclease SbcC
GI-48	2	VCMZO3.2627~2637	11	VCA0569-VCA0570		Vibrio f237 like prophage
GI-49	1	VCMZO3.3722~3740	19	VC1018-VC1019		T7 like prophage
GI-50	1	VCMZO3.3434~3441/3185~3188	12	VC0174-VC0186		Putative prophage
GI-51	2	VCAM19226.3195~3223	29	VCA0281-VCA0283		Putative prophage
GI-52	1	VCAM19226.1624~1629	6	VC0209-VC0208		Bacteriocin/antobiotic exporter
GI-53	1	VC1587.1591~1658	68	VC1710?-VC1711		Putative prophage
GI-54	1	VC1587.1521/1701~1713	32-40	VC1757-VC1773		Type III secretion
GI-55	1	VC1587.841~848	8	VC1340		Integrase/ DNA ligase
GI-56	1	VC62339.3453~3465	13	VC0489-VC0517		Iintegrase/DNA repair/helicase/hipA protein
GI-57	2	VC62339.1537~1541	2-5	VCA0022-VCA0018		hypothetical proteins
GI-58	2	VC62339.1772~1778	7	VCA0235-VCA0237		Mu-like prophage
GI-59	1	VC62339.2470~2488	19	VC1411-VC1414?		TETRATRICOPEPTIDE REPEAT FAMILY PROTEIN/Cell division protein ftsH
GI-60	1	VCV51.1633~1679/2661~2677	64	VC1757-VC1758		Putative prophage
GI-61	2	VCV51.2939~2943	5	VCA1067-VCA1064		Outer membrane lipoprotein-sorting protein/Predicted exporter of the RND superfamily/Transcriptional activator, LuxR/UhpA family
GI-62	1	VCVL426.40~44	5	VC1414-VC1407		Type I restriction-modification system
GI-63	1	VCVL426.142~146	5	VC1519-VC1520		DNA helicase
GI-64	1	VCVL426.187~197	11-13	VC1560-VC1563		Lux operon
GI-65	1	VCVL426.786~793	8	VC2231-VC2232		Arylsulfatase
GI-66	1	VCVL426.2052~2057	6	VC0806-VC0807		Ferric iron ABC transporter/ iron-binding protein
GI-67	2	VCVL426.3389~3405	17	VCA0735-VCA0734		Choline dehydrogenase/Betaine aldehyde dehydrogenase/iron aquisition yersiniabactin synthesis enzyme/FyuA precursor
GI-68	1	VCRC385.1973~1991	19	VC0209-VC0208		Transposable element

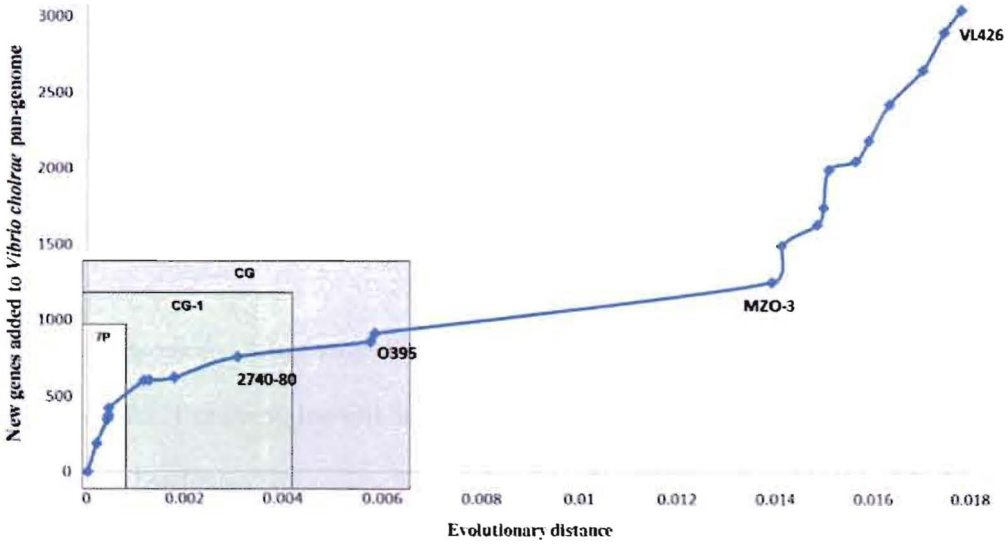
GI-69	1	VCRC385.2016~2024/3030~3032	12	VC0168-VC0167		Flp pilus assembly/ Type II/IV secretion system
GI-70	1	VCRC385.2972~2976	5	VC2715		Possible MFS Superfamliy multidrug-efflux transporter;MccB-like protein
GI-71	1	VCRC385.3237~3250	14	VC0081-VC0080		Type I restriction-modification system/ Tn7-like transposition proteins
GI-72	1	VCRC385.96~118/1120	24	VC1407-VC1414		Metabolism of amino acids and derivatives/carbohydrates
GI-73	1	VCRC385.490~494	5	VC1669-VC1667?		Chalcone synthase/Cytotoxic necrotizing factor 1

ORFs of one representative genome were given for a genomic island with similar gene content, organization and sequence similarity among test strains. If significant variation is found for single genomic island class, they were further divided into subgroups (a,b,c,d) based on gene contents. Chromosome numbers, i.e. 1(large) or 2(small), were given where they are present both chromosomes.

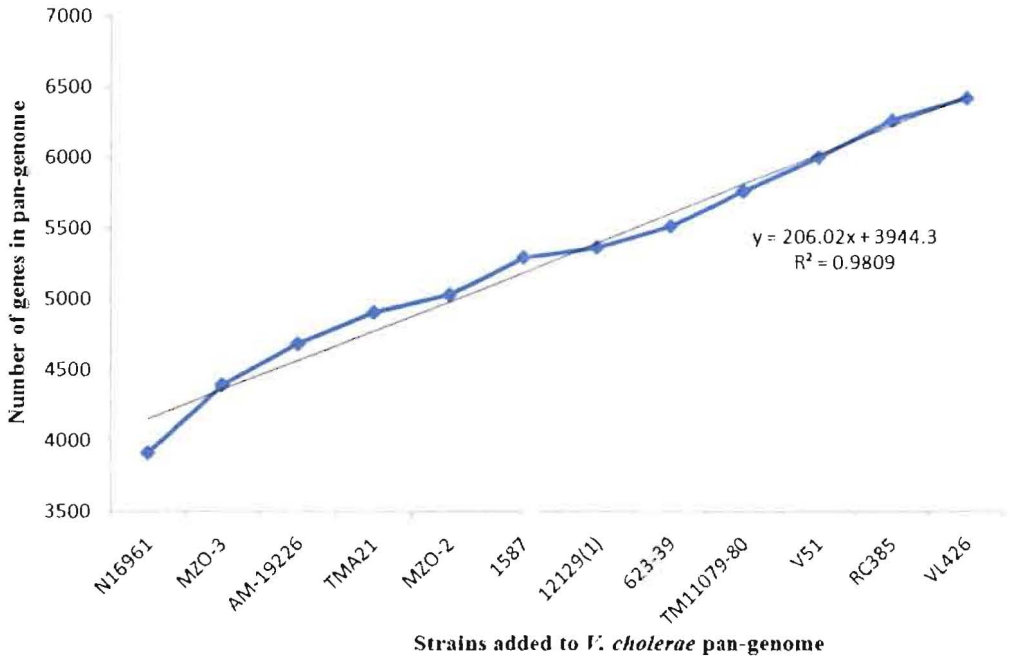
Details on *V. cholerae* ORFs were given in Supplementary Table 1.


Evolutionary Distance
Strains

(a)

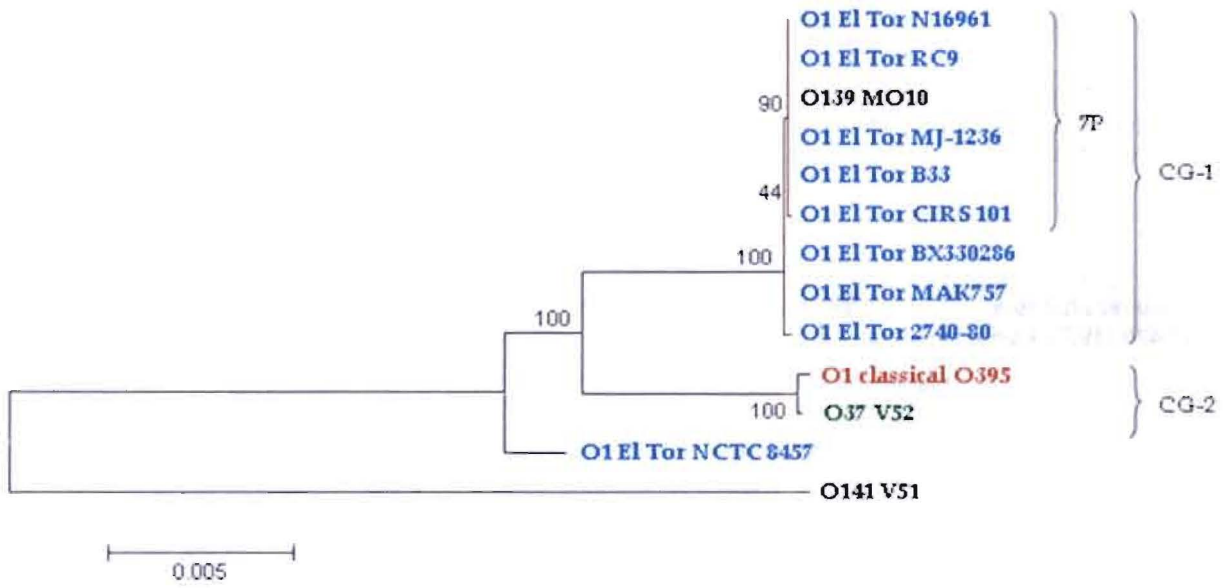

New genes added to Vibrio cholrae pan-genome
Evolutionary distance
CG
CG-1
7P
2740-80
O395
MZO-3
VL426
3000
2500
2000
1500
1000
500
0
0
0.002
0.004
0.006
0.008
0.01
0.012
0.014
0.016
0.018

(b)

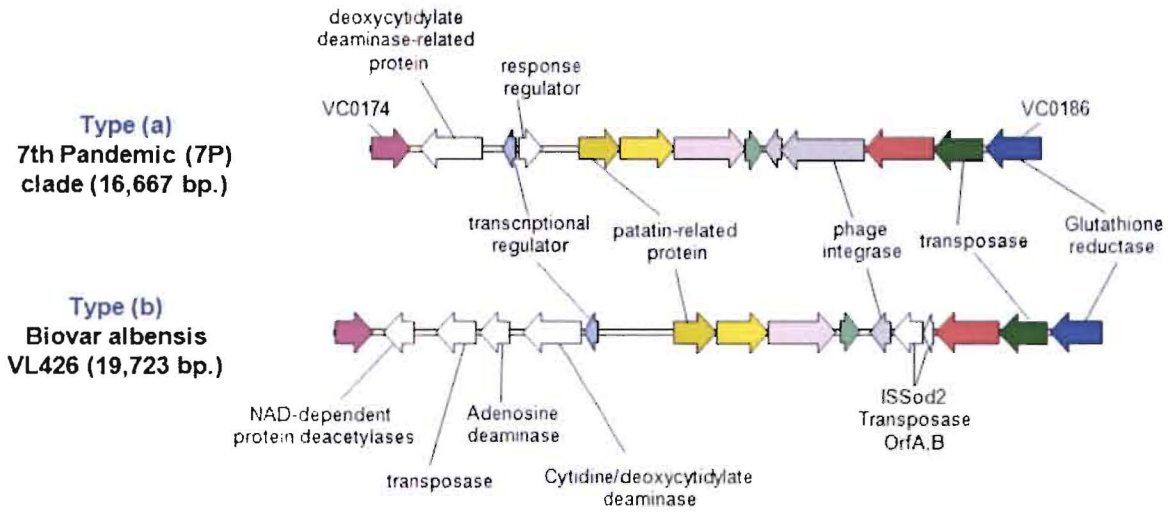


(c)

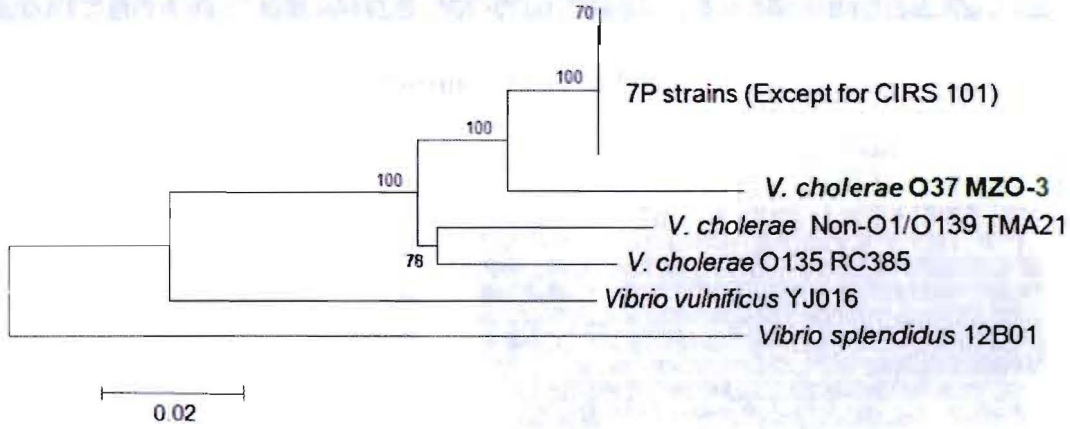
Supplementary Figure 1. Pan-genome analysis of *Vibrio cholerae*. (a) Evolutionary distances of *V. cholerae* strains from the reference strain (O1 El Tor N16961). (b) cumulative number of newly found additional genes to *V. cholerae* pan-genome with N16961 as reference. Strains were added in the order as shown in (a). (c) Pan-genome calculated from strains that are representatives of 12 distinct phyletic lines (Fig. 1a). Linear regression indicates that 206 genes will be added to the pan-genome of *V. cholerae*, if new strain represents equally distinct phyletic lineage.



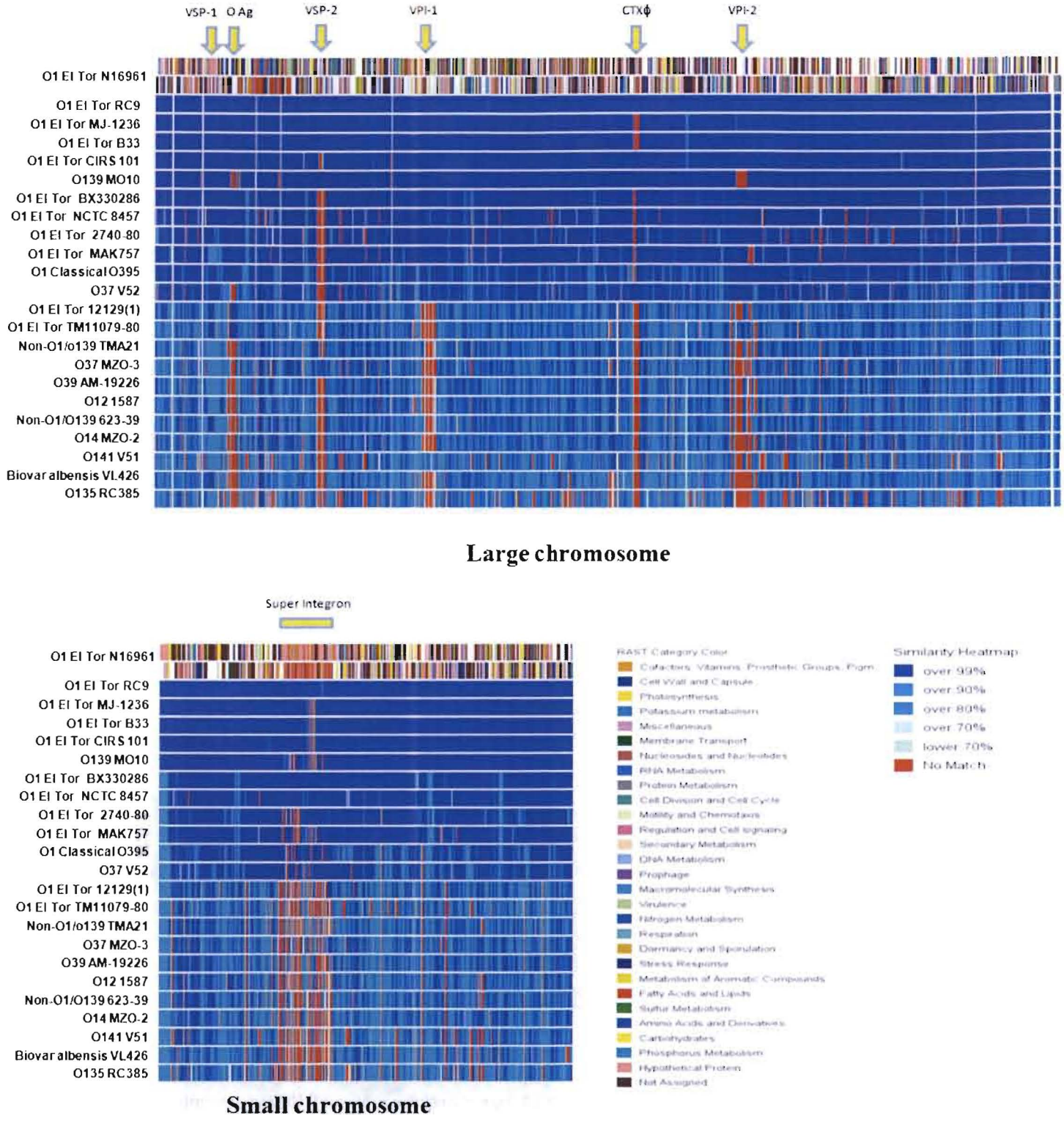
Supplementary Figure 2. Neighbor-joining tree showing evolutionary relationships of *Vibrio* pathogenicity island (VPI-1) found in *V. cholerae* strains. The calculation was based on 24 orthologous ORFs (VC0821-VC0834,VC0836-VC0845) comprising 27,393 bp. The phylogenetic placement of *V. cholerae* O1 El Tor NCTC 8457 (CG-1 member) indicates the replacement of its VPI-1 by that from a non-CG strain. Bar represents 0.005 substitution per site.



Supplementary Figure 3. Genetic organization of *Vibrio* seventh pathogenicity island-1 (VSP-1) genomic island (N16961 ORFs: VC0175-VC0185) found in only 7th pandemic (7P) *V. cholerae* strains and *V. cholerae* biovar albensis VL426.



Supplementary Figure 4. Neighbor-joining tree showing evolutionary relationships of *Vibrio* seventh pathogenicity island (VSP-2) found in strains of *V. cholerae* and related taxa. The VSP-2 like genomic islands may be wide spread among vibrios. The calculation was based on eleven orthologous ORFs (4840 bp long; VC0495-VC0498, VC0504-VC0510). Bar represents 0.02 substitution per site.



Supplementary Figure 5. Gene contents of 22 *V. cholerae* strains showing gene conservation using *V. cholerae* O1 El Tor N16961 as the reference strain. Orthologues found are color-coded according to nucleotide sequence similarity values to the reference ORFs. Major genomic islands are indicated at the top row.

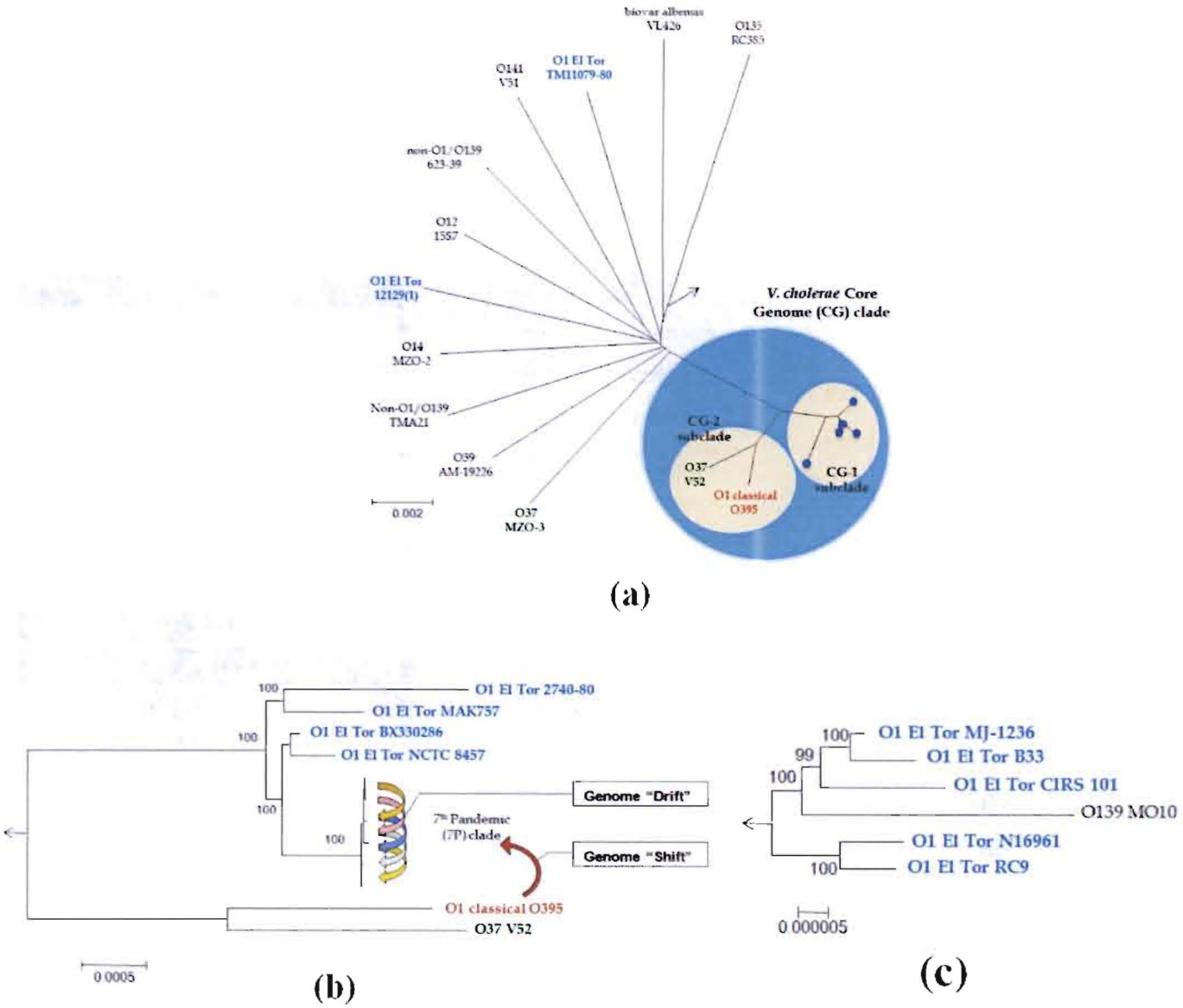
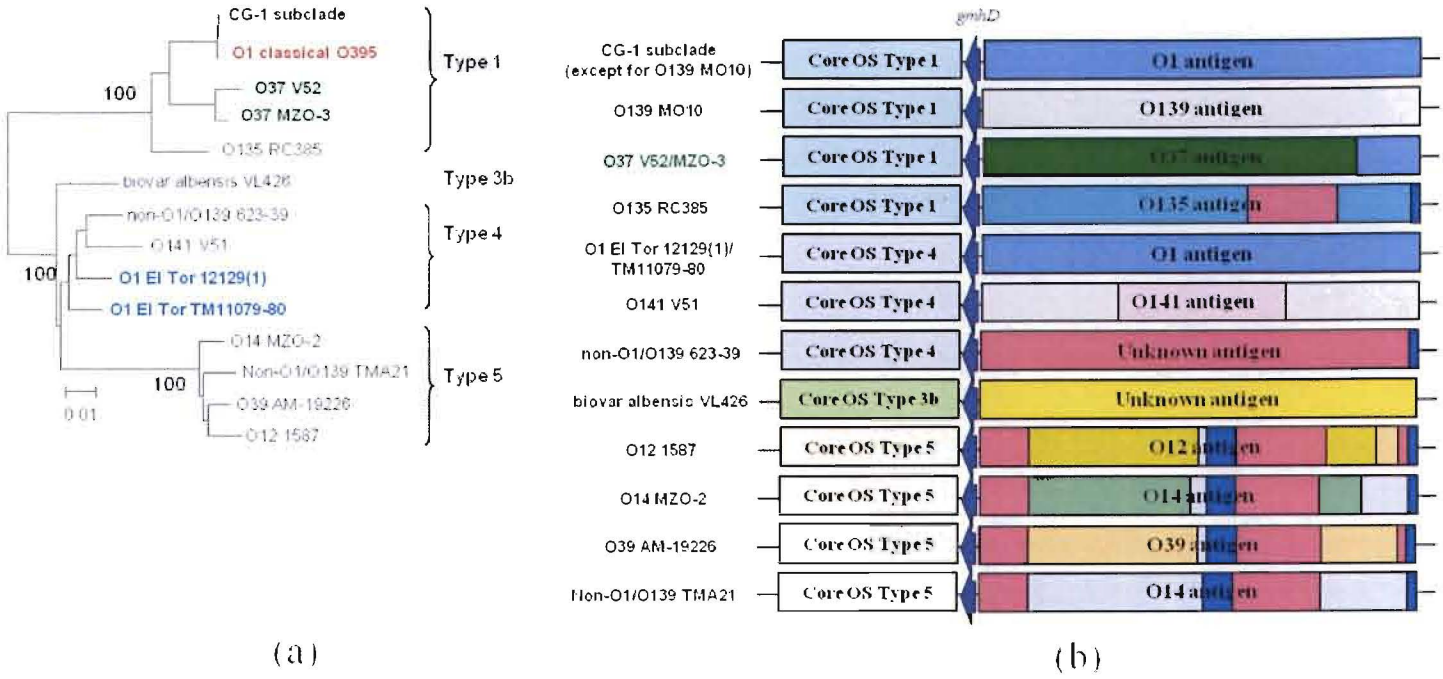
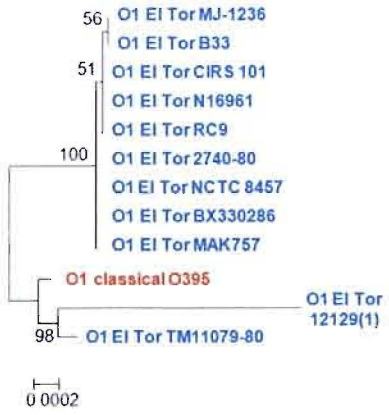

biovar albensis
VL426
O135
RC385
O141
V51
O1 El Tor
TM11079-80
non-O1/O139
623-39
O12
1557
O1 El Tor
12129(1)
O14
MZO-2
Non-O1/O139
TMA21
O39
AM-19226
O37
MZO-3
0.002
V. cholerae Core
Genome (CG) clade
CG-2
subclade
CG-1
subclade
O37
V52
O1 classical
O395
(a)
100
O1 El Tor 2740-80
O1 El Tor MAK757
100
O1 El Tor BX330286
O1 El Tor NCTC 8457
100
Genome "Drift"
7th Pandemic
(7P) clade
100
Genome "Shift"
O1 classical O395
O37 V52
0.0005
(b)
100
O1 El Tor MJ-1236
99
O1 El Tor B33
100
O1 El Tor CIRS 101
O139 MO10
O1 El Tor N16961
100
O1 El Tor RC9
0.000005
(c)

Figure 1.


CG-1 subclade
O1 classical O395
O37 V52
O37 MZO-3
O135 RC385
biovar albensis VL426
non-O1/O139 623-39
O141 V51
O1 El Tor 12129(1)
O1 El Tor TM11079-80
O14 MZO-2
Non-O1/O139 TMA21
O39 AM-19226
O12 1587
100
100
100
0 01
Type 1
Type 3b
Type 4
Type 5
gmhD
CG-1 subclade (except for O139 MO10)
O139 MO10
O37 V52/MZO-3
O135 RC385
O1 El Tor 12129(1)/ TM11079-80
O141 V51
non-O1/O139 623-39
biovar albensis VL426
O12 1587
O14 MZO-2
O39 AM-19226
Non-O1/O139 TMA21
Core OS Type 1
Core OS Type 1
Core OS Type 1
Core OS Type 1
Core OS Type 4
Core OS Type 4
Core OS Type 4
Core OS Type 3b
Core OS Type 5
Core OS Type 5
Core OS Type 5
Core OS Type 5
O1 antigen
O139 antigen
O135 antigen
O1 antigen
O141 antigen
Unknown antigen
Unknown antigen

(a) (b)


56
51
100
98
O1 El Tor MJ-1236
O1 El Tor B33
O1 El Tor CIRS 101
O1 El Tor N16961
O1 El Tor RC9
O1 El Tor 2740-80
O1 El Tor NCTC 8457
O1 El Tor BX330286
O1 El Tor MAK757
O1 classical O395
O1 El Tor 12129(1)
O1 El Tor TM11079-80
0 0002

(c)

Figure 2.

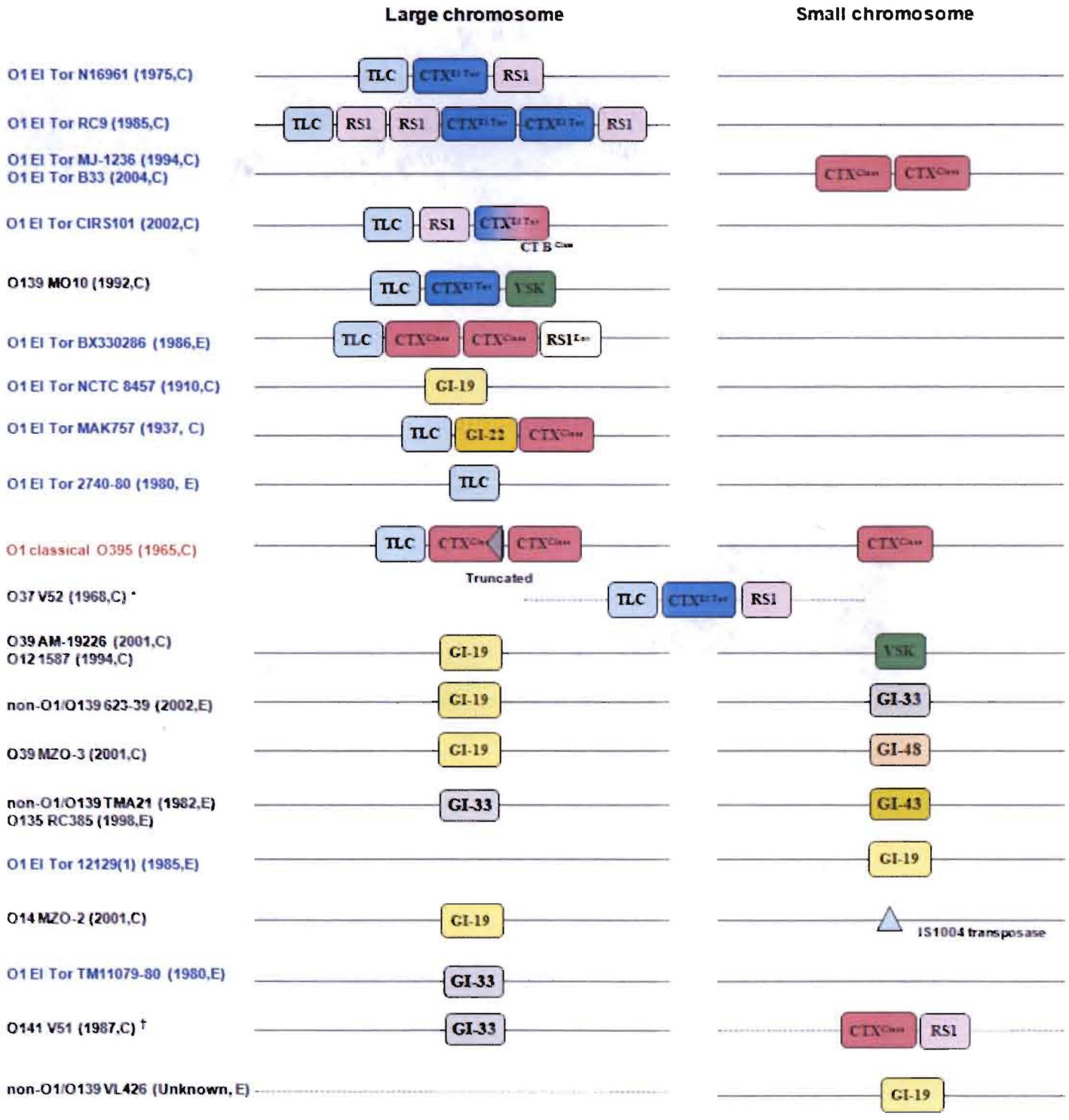

Large chromosome
Small chromosome
O1 El Tor N16961 (1975,C)
O1 El Tor RC9 (1985,C)
O1 El Tor MJ-1236 (1994,C)
O1 El Tor B33 (2004,C)
O1 El Tor CIRS101 (2002,C)
O139 MO10 (1992,C)
O1 El Tor BX330286 (1986,E)
O1 El Tor NCTC 8457 (1910,C)
O1 El Tor MAK757 (1937, C)
O1 El Tor 2740-80 (1980, E)
O1 classical O395 (1965,C)
O37 V52 (1968,C) *
O39 AM-19226 (2001,C)
O12 1587 (1994,C)
non-O1/O139 623-39 (2002,E)
O39 MZO-3 (2001,C)
non-O1/O139 TMA21 (1982,E)
O135 RC385 (1998,E)
O1 El Tor 12129(1) (1985,E)
O14 MZO-2 (2001,C)
O1 El Tor TM11079-80 (1980,E)
O141 V51 (1987,C) †
non-O1/O139 VL426 (Unknown, E)
TLC
RS1
VSK
GI-19
GI-22
GI-33
GI-43
GI-48
Truncated
IS1004 transposase

Figure 3.

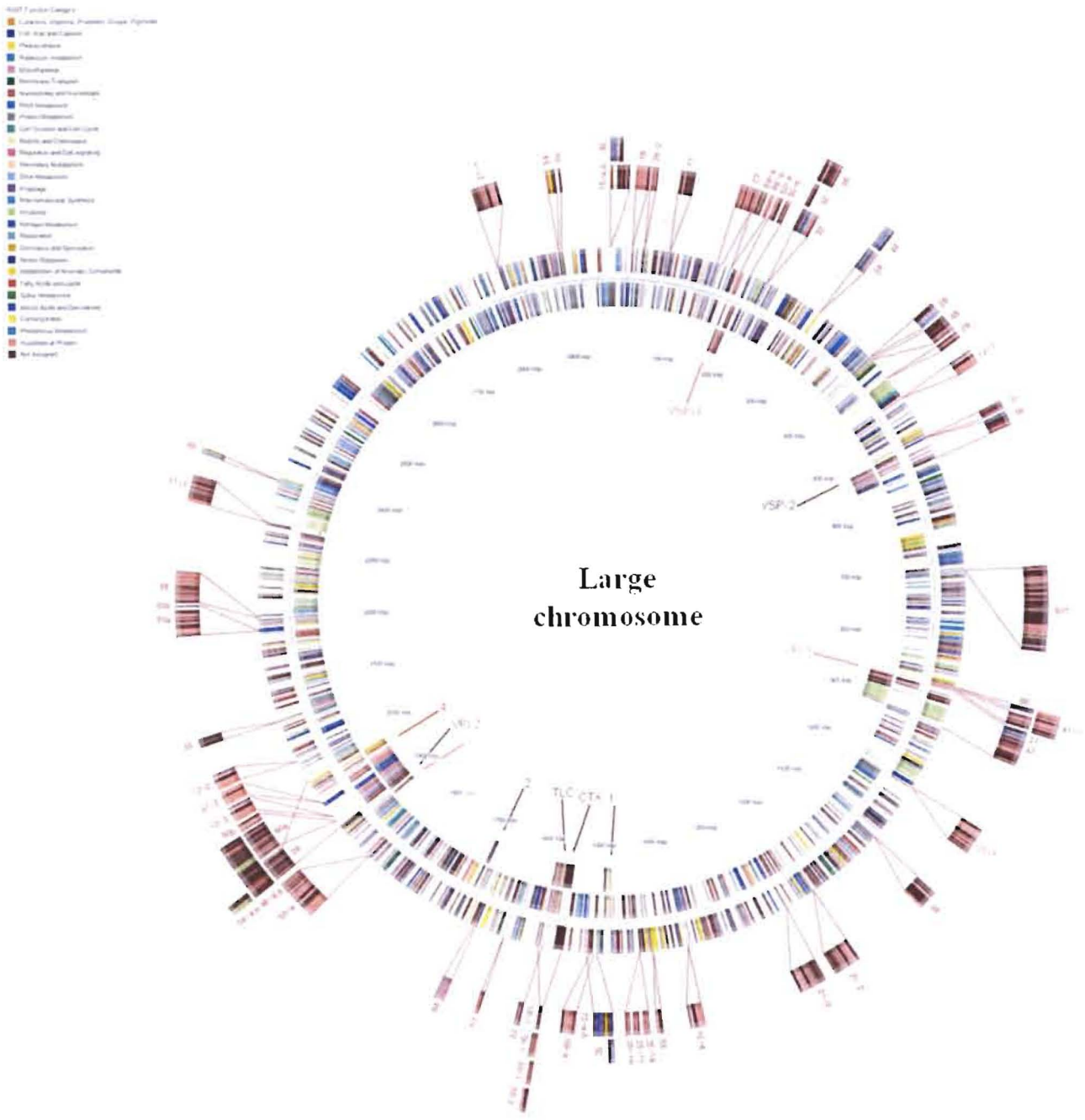


Figure 4 .

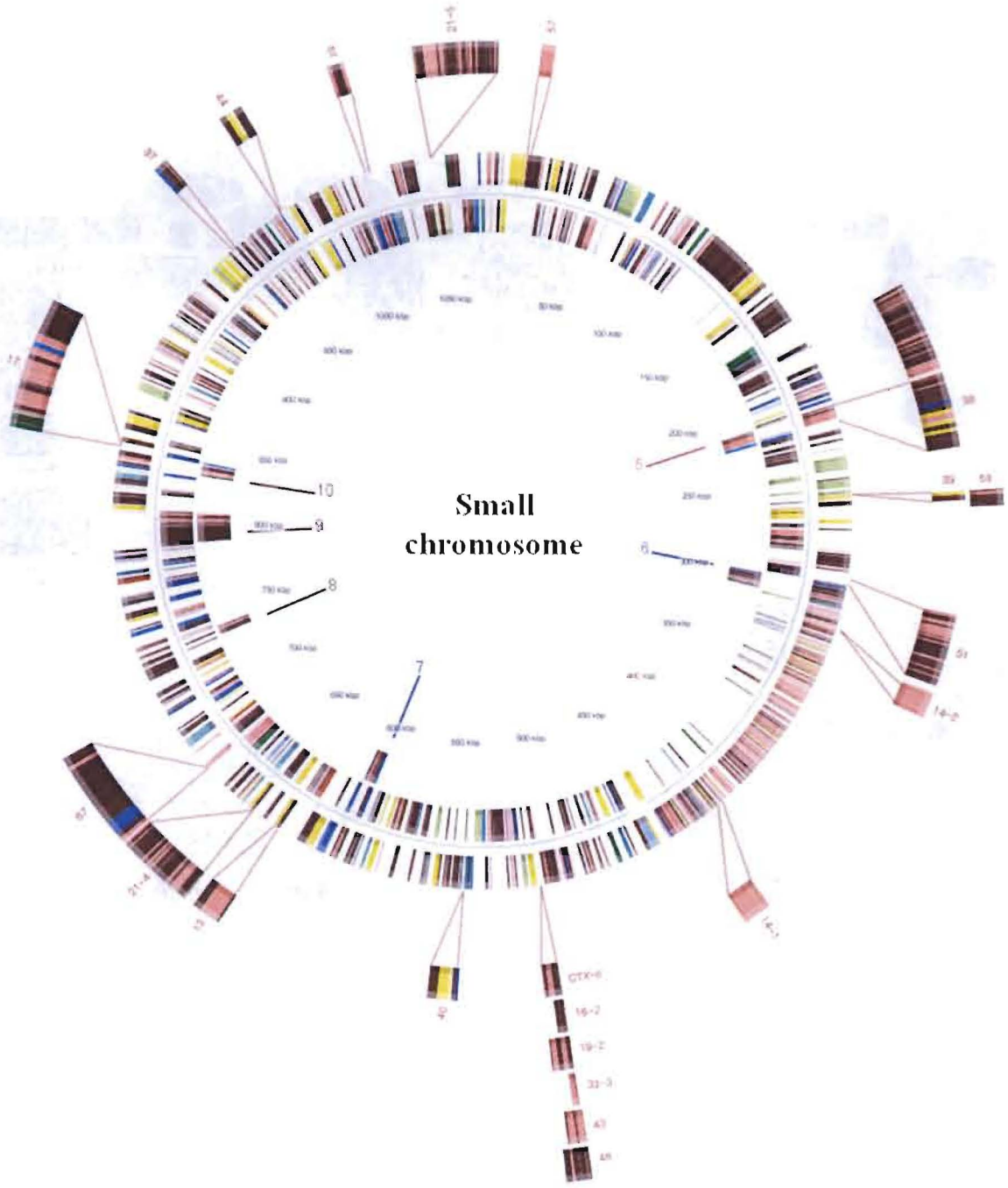


Figure 4 .

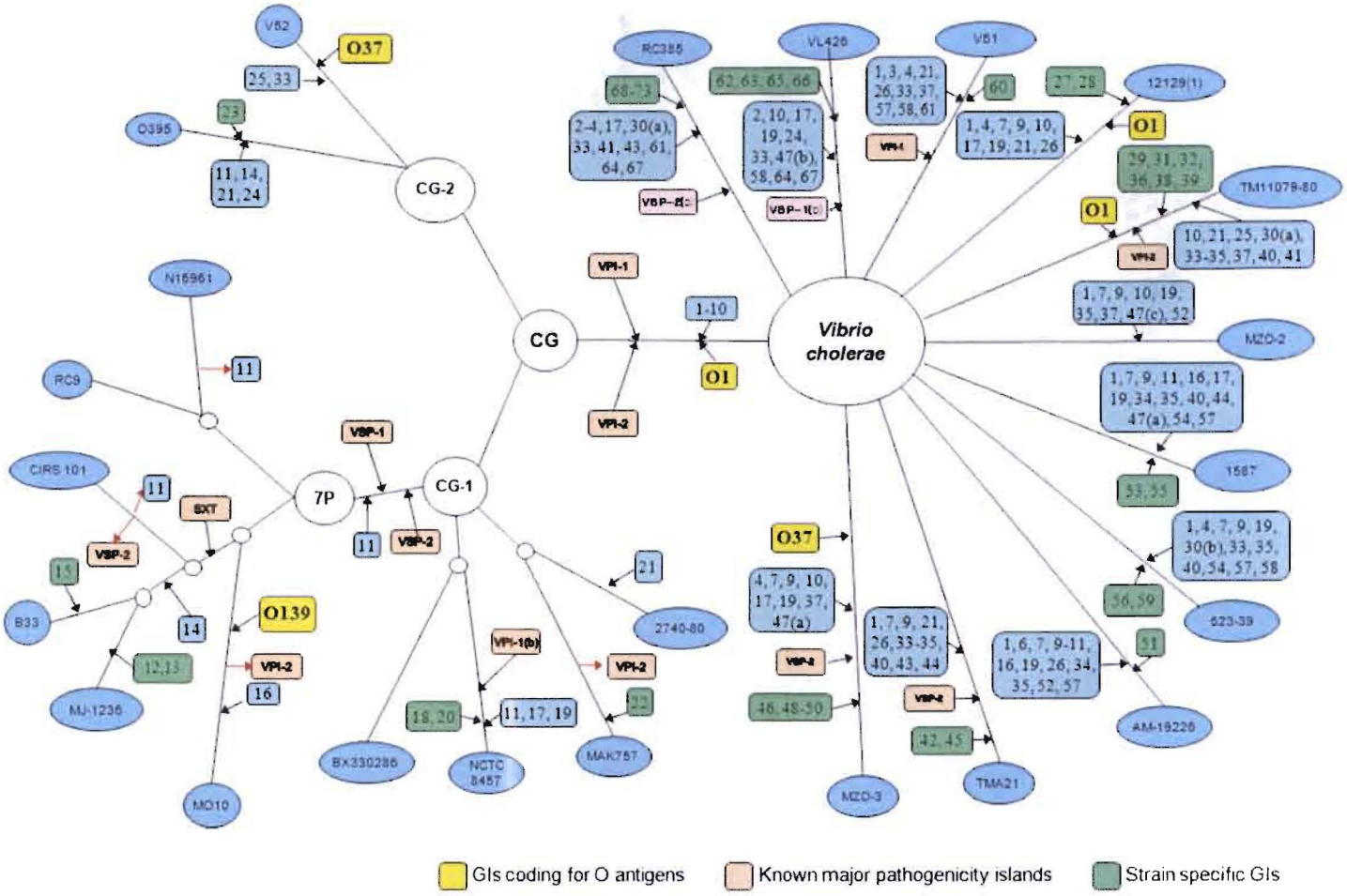

Vibrio cholerae
CG
CG-2
CG-1
7P
O1
O37
O139
V52
25,33
23
O395
11,14,
21,24
N16961
11
RC9
CIRS 101
SXT
VSP-2
15
B33
14
12,13
MJ-1236
16
MO10
VSP-1
11
BX330286
18,20
11,17,19
NCTC 8457
VPI-1(b)
MAK757
22
VPI-2
21
2740-80
RC385
68-73
2-4,17,30(a),
33,41,43,61,
64,67
VPI-1
1-10
VPI-2
VL426
62,63,65,66
2,10,17,
19,24,
33,47(b),
58,64,67
V51
1,3,4,21,
26,33,37
57,58,61
60
1,4,7,9,10,
17,19,21,26
27,28
12129(1)
29,31,32,
36,38,39
TM11079-80
10,21,25,30(a),
33-35,37,40,41
1,7,9,10,19,
35,37,47(c),52
MZO-2
1,7,9,11,16,17,
19,34,35,40,44,
47(a),54,57
1587
53,55
1,4,7,9,19,
30(b),33,35,
40,54,57,58
56,59
623-39
51
AM-19226
1,6,7,9-11,
16,19,26,34,
35,52,57
4,7,9,10,
17,19,37,
47(a)
1,7,9,21,
26,33-35,
40,43,44
46,48-50
42,45
MZO-3
TMA21
GIs coding for O antigens
Known major pathogenicity islands
Strain specific GIs

Figure 5.

Table 1. Characteristics of Vibrio cholerae strains compared in this study.

Strain	Genome Code	Serogroup	Biotype	Geographical Origin	Source of isolation	Year of isolation	Sequencing status[a]	Number of contigs	Accession
N16961	VCN16961	O1 Inaba	El Tor	Bangladesh	Clinical	1975	Complete	2	NC_002505/ NC_002506
RC9	VCRC9	O1 Ogawa	El Tor	Kenya	Clinical	1985	S/4/E	11	This study
MJ-1236	VCMJ1236	O1 Inaba	El Tor	Matlab, Bangladesh	Clinical	1994	Complete	2	This study
B33	VCB33	O1 Ogawa	El Tor	Beira, Mozambique	Clinical	2004	S/4/E	17	This study
CIRS 101	VCCIRS101	O1 Inaba	El Tor	Dhaka, Bangladesh	Clinical	2002	S/4/E	18	This study
MO10	VCMO10	O139		Madras,India	Clinical	1992	S/4	84	AAKF03000000
2740-80	VC274080	O1 Inaba	El Tor	US Gulf Coast	Water	1980	Sanger	257	AAUT01000000
BX330286	VCBX330286	O1 Inaba	El Tor	Australia	Water	1986	Complete	8	This study
MAK757	VCMAK757	O1 Ogawa	El Tor	Celebes Islands	Clinical	1937	Sanger	206	AAUS00000000
NCTC 8457	VC8457	O1 Inaba	El Tor	Saudi Arabia	Clinical	1910	Sanger	390	AAWD01000000
O395	VCO395	O1 Ogawa	Classical	India	Clinical	1965	Complete	2	NC_009456/ NC_009457
V52	VCV52	O37		Sudan	Clinical	1968	Sanger	268	AAKJ02000000
12129(1)	VC12129	O1 Inaba	El Tor	Australia	Water	1985	S/4/E	12	This study
TM 11079-80	VCTM11079	O1 Ogawa	El Tor	Brazil	Sewage	1980	S/4/E	35	This study
VL426	VCVL426	non-O1/O139	albensis	Maidstone, Kent, UK	Water	Unknown	Complete	5	This study
TMA21	VCTMA21	non-O1/O139		Brazil	Seawater	1982	S/4/E	20	This study
1587	VC1587	O12		Lima, Peru	Clinical	1994	Sanger	254	AAUR01000000
RC385	VCRC385	O135		Chesapeake Bay	Plankton	1998	Sanger	550	AAKH02000000
MZO-2	VCMZO2	O14		Bangladesh	Clinical	2001	Sanger	162	AAWF01000000
V51	VCV51	O141		USA	Clinical	1987	Sanger	360	AAKI02000000
MZO-3	VCMZO3	O37		Bangladesh	Clinical	2001	Sanger	292	AAUU01000000
AM-19226	VCAM19226	O39		Bangladesh	Clinical	2001	Sanger	154	AATY01000000
623-39	VC62339	non-O1/O139		Bangladesh	Water	2002	Sanger	314	AAWG00000000

ger, Draft assemblies by Sanger sequencing; S/4, Sanger sequencing and 454 pyrosequencing were combined; , S/4 followed by quality improvement by standard genome sequencing procedures.

Supplementary Table 3. *Vibrio cholerae* genomic islands that occupy the same position/area, showing a cassette like property.

	Insertion region	GI-designation	Strains
Large chromosome	VC0174-VC0186	VSP-1	N16961, RC9, MJ-1236, B33, MO10
		GI-50	MZO-3
	VC0489 -VC0517	VSP-2	N16961, RC9, B33, MJ1236, MO10, TMA21, MZO-3
		GI-56	623-39
	VC0809 - VC0848	VPI-1	N16961, RC9, B33, MJ1236, MO10, BX330286, NCTC 8457, 2740-80, MAK757, O395, V52, V51
		GI-47	MZO-3, 1587, MZO-2, VL426
	VC1757 - VC1810	VPI-2	N16961, RC9, MJ1236, B33, MO10, BX330286, NCTC8457, 2740-80, MAK757, O395, V52, TM11079-80
		GI-26	12129(1), TMA21, AM-19226, V51
		GI-46	MZO-3
		GI-54	1587, 623-39
	VC1451 - VC1465	CTX□	N16961, RC9, O395
		GI-16	MO10
		GI-19	NCTC 8457, AM-19226, 1587, 623-39, MZO-2, MZO-3
		GI-22	MAK757
		GI-33	TM11079-80, TMA21, V51, RC385
	VC0002 - VC0003	GI-15	B33
		GI-30	TM11079-80, 623-39, RC385
	VC0289 - VC0290	GI-24	O395, VL426
		GI-42	TMA21
	VC0209 - VC0208	GI-32	TM11079-80
		GI-52	AM-19226, MZO-2
		GI-68	RC385
	VC0847 - VC0807	GI-27	12129(1)
		GI-41	TM11079-80
	VC1407-VC1414	GI-59	623-39
		GI-62	VL426
		GI-72	RC385
Small chromosome	VCA0569 - VCA0570	CTX□	MJ-1236, B33, O395
		GI-16	AM-19226, 1587
		GI-19	12129(1),VL426
		GI-33	623-39
		GI-43	TMA21, RC385
		GI-48	MZO-3
	VCA0197 - VCA0204	GI-5	N16961, RC9, B33, MJ1236, MO10, BX330286, NCTC8457, 2740-80, MAK757, O395, V52
		GI-38	TM11079-80
	VCA0235 - VCA0237	GI-39	TM11079-80
		GI-58	623-39, V51, VL426