Genetic Diversity of Toxigenic and Nontoxigenic Vibrio cholerae Serogroups O1 and O139 Revealed by Array-Based Comparative Genomic Hybridization

Bo Pang,1† Meiying Yan,1† Zhigang Cui,1 Xiaofen Ye,2 Baowei Diao,1 Yonghong Ren,2 Shouyi Gao,1 Liang Zhang,2 and Biao Kan1*

State Key Laboratory for Infectious Disease Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, P.O. Box 5, Chaoephing, Beijing 102206, People’s Republic of China,1 and National Engineering Research Center for Beijing Biochip Technology, 18 Life Science Parkway, Beijing 102206, People’s Republic of China2

Received 25 December 2006/Accepted 21 April 2007

Toxigenic serogroups O1 and O139 of Vibrio cholerae may cause cholera epidemics or pandemics. Nontoxigenic strains within these serogroups also exist in the environment, and also some may cause sporadic cases of disease. Herein, we investigate the genomic diversity among toxigenic and nontoxigenic O1 and O139 strains by comparative genomic microarray hybridization with the genome of El Tor strain N16961 as a base. Conservation of the toxigenic O1 El Tor and O139 strains is found as previously reported, whereas accumulation of genome changes was documented in toxigenic El Tor strains isolated within the 40 years of the seventh pandemic. High phylogenetic diversity in nontoxigenic O1 and O139 strains is observed, and most of the genes absent from nontoxigenic strains are clustered together in the N16961 genome. By comparing these toxigenic and nontoxigenic strains, we observed that the small chromosome of V. cholerae is quite conservative and stable, outside of the superintegron region. In contrast to the general stability of the genome, the superintegron demonstrates pronounced divergence among toxigenic and nontoxigenic strains. Additionally, sequence variation in virulence-related genes is found in nontoxigenic El Tor strains, and we speculate that these intermediate strains may have pathogenic potential should they acquire CTX prophage alleles and other gene clusters. This genome-wide comparison of toxigenic and nontoxigenic V. cholerae strains may promote understanding of clonal differentiation of V. cholerae and contribute to an understanding of the origins and clonal selection of epidemic strains.

Vibrio cholerae is a natural inhabitant in warm aquatic ecosystems (13). No firm understanding is available regarding how many pandemics have been caused by V. cholerae in the course of human events; however, it is generally accepted that at least seven distinct cholera pandemics have occurred since 1817 (2).

Of the more than 200 O-antigen serogroups of V. cholerae identified, only O1 and O139 are known to cause epidemic and pandemic cholera. The fifth and sixth pandemics (caused by a classical biotype of the O1 serogroup) were confirmed etiologically, and in 1961 the seventh and present pandemic began, caused by the O1 El Tor biotype (2). In 1992, a novel serogroup, O139, emerged and caused epidemic cholera in Bangladesh and India (1, 5); the O139 serogroup remains confined to Southeast Asia (49).

Cholera toxin is the major virulence factor of toxigenic V. cholerae (31), and colonization of the intestine is the first major step in Vibrio pathogenesis. The lysogenic infection of nontoxigenic V. cholerae strains by CTXΦ, which carries ctxAB, is an important mechanism for the emergence of new toxigenic strains (54). The toxin-coregulated pilus (TCP) is the receptor for CTXΦ and plays an important role in intestinal colonization by V. cholerae. The TCP is located in the V. cholerae pathogenicity island (VPI). Other genomic islands, including Vibrio seventh-pandemic island I (VSP-I) and VSP-II, are present in toxigenic O139 and El Tor strains of the seventh pandemic (16).

V. cholerae can coexist with a variety of zooplankton and phytoplankton (29), and such survival in aquatic systems is regarded as a prerequisite for cholera pandemics. A report showed that strains with virulence potential (from among the majority of genetically diverse nonpathogenic strains) can be enriched by passage through mammalian host intestines (20).

Analysis of diverse V. cholerae strains isolated from different sources has revealed that the strains that can cause diarrhea are highly clonal and that they have evolved independently from different lineages of nontoxigenic environmental isolates. Additionally, it is known that the various nontoxigenic environmental lineages are composed of heterogeneous populations (20, 32, 50). Several studies have indicated that virulence genes disperse among different serogroups of nontoxigenic strains that are present in aquatic ecosystems and that these aquatic strains can be virulence gene reservoirs for toxigenic strains (20, 21, 42).

In order to understand the origins of toxigenic strains, the investigation of genomic differences among V. cholerae strains (especially nontoxigenic strains) will provide insight. The genome contents of toxigenic and nontoxigenic strains have been studied by genome sequencing or comparative genome hybridization.
non-China exception that was isolated in India. For nontoxigenic O1 and O139 pandemic (from 1961 to 2005). Serogroup O139 strains were isolated in different isolated from three cholera epidemics in China during the seventh cholera strains isolated in Bangladesh and India, toxigenic O1 El Tor strains were may provide further insight into the origins of the seventh-

Speculation that the nontoxigenic strains carrying different alleles may represent different intermediate strains in the orig-

Materials and Methods

Strains. The strains used in this study are listed in Table 1. Except for the strains isolated in Bangladesh and India, toxigenic O1 El Tor strains were isolated from three cholera epidemics in China during the seventh cholera pandemic (from 1961 to 2005). Serogroup O139 strains were isolated in different Chinese regions and in different years since their emergence in 1992. MO45 is a non-China exception that was isolated in India. For nontoxigenic O1 and O139 strains, strains with extensive differences in source and year of isolation were selected. Four nontoxigenic El Tor strains containing TCP clusters were also included in this study.

Chromosomal DNA preparation. All strains were grown in 5 ml Luria-Bertani broth to an optical density at 600 nm of 0.8. Chromosomal DNA was prepared with NucleoSpin Tissue kits (Macherey-Nagel, Germany) according to the man-

Construction of microarrays. DNA microarrays were constructed on the basis of the genomic content of N16961 (GenBank accession no. NC002505 and NC002506). An in-house high-throughput computer algorithm was used to de-

TABLE 1. Strains used for comparative analysis

<table>
<thead>
<tr>
<th>Strain</th>
<th>Serogroup, biotype</th>
<th>Serotype</th>
<th>Yr isolated</th>
<th>Where isolated</th>
<th>Origin</th>
<th>ctcAB</th>
<th>No. of genes absent</th>
<th>No. of genes absent in superintegron</th>
</tr>
</thead>
<tbody>
<tr>
<td>V0518</td>
<td>O1, El Tor</td>
<td>Ogawa</td>
<td>2005</td>
<td>China</td>
<td>Shrimp</td>
<td>199</td>
<td>51</td>
<td></td>
</tr>
<tr>
<td>SD5001</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>1995</td>
<td>China</td>
<td>Water</td>
<td>203</td>
<td>60</td>
<td></td>
</tr>
<tr>
<td>93097</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>1993</td>
<td>China</td>
<td>Patient</td>
<td>199</td>
<td>52</td>
<td></td>
</tr>
<tr>
<td>V0550</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>2005</td>
<td>China</td>
<td>Patient</td>
<td>309</td>
<td>66</td>
<td></td>
</tr>
<tr>
<td>19-22</td>
<td>O1, El Tor</td>
<td>Ogawa</td>
<td>1977</td>
<td>China</td>
<td>Patient</td>
<td>244</td>
<td>17</td>
<td></td>
</tr>
<tr>
<td>JS32</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>1990</td>
<td>China</td>
<td>Patient</td>
<td>309</td>
<td>111</td>
<td></td>
</tr>
<tr>
<td>7743</td>
<td>O1, El Tor</td>
<td>Ogawa</td>
<td>1977</td>
<td>China</td>
<td>Water</td>
<td>112</td>
<td>28</td>
<td></td>
</tr>
<tr>
<td>FJ147</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>2005</td>
<td>China</td>
<td>Patient</td>
<td>34</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>ZJ0596</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>2005</td>
<td>China</td>
<td>Patient</td>
<td>34</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>V011506</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>2001</td>
<td>China</td>
<td>Patient</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>200106</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>2001</td>
<td>China</td>
<td>Water</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>93284</td>
<td>O1, El Tor</td>
<td>Ogawa</td>
<td>1993</td>
<td>China</td>
<td>Patient</td>
<td>16</td>
<td>16</td>
<td></td>
</tr>
<tr>
<td>FJ0004</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>1980</td>
<td>China</td>
<td>Patient</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>D118</td>
<td>O1, El Tor</td>
<td>Ogawa</td>
<td>1961</td>
<td>China</td>
<td>Patient</td>
<td>27</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>6312</td>
<td>O1, El Tor</td>
<td>Ogawa</td>
<td>1961</td>
<td>Indonesia</td>
<td>Patient</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>N16961</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>1975</td>
<td>Bangladesh</td>
<td>Patient</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>79005</td>
<td>O1, El Tor</td>
<td>Ogawa</td>
<td>1979</td>
<td>China</td>
<td>Patient</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>Wujiang-2</td>
<td>O1, El Tor</td>
<td>Inaba</td>
<td>1980</td>
<td>China</td>
<td>Water</td>
<td>71</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>FJ299957</td>
<td>O139</td>
<td></td>
<td>1999</td>
<td>China</td>
<td>Patient</td>
<td>0</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>93101</td>
<td>O139</td>
<td></td>
<td>1993</td>
<td>China</td>
<td>Patient</td>
<td>72</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>93209</td>
<td>O139</td>
<td></td>
<td>1993</td>
<td>China</td>
<td>Carrier</td>
<td>74</td>
<td>22</td>
<td></td>
</tr>
<tr>
<td>MO45</td>
<td>O139</td>
<td></td>
<td>1993</td>
<td>India</td>
<td>Patient</td>
<td>71</td>
<td>21</td>
<td></td>
</tr>
<tr>
<td>98106</td>
<td>O139</td>
<td></td>
<td>1998</td>
<td>China</td>
<td>Patient</td>
<td>72</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>200306</td>
<td>O139</td>
<td></td>
<td>2003</td>
<td>China</td>
<td>Patient</td>
<td>73</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>SCH40482</td>
<td>O139</td>
<td></td>
<td>2004</td>
<td>China</td>
<td>Water</td>
<td>592</td>
<td>50</td>
<td></td>
</tr>
<tr>
<td>B4</td>
<td>O139</td>
<td></td>
<td>1994</td>
<td>China</td>
<td>Patient</td>
<td>599</td>
<td>23</td>
<td></td>
</tr>
<tr>
<td>94001</td>
<td>O139</td>
<td></td>
<td>1994</td>
<td>China</td>
<td>Water</td>
<td>405</td>
<td>105</td>
<td></td>
</tr>
<tr>
<td>569B</td>
<td>O1, classical</td>
<td>Inaba</td>
<td>1948</td>
<td>India</td>
<td>Patient</td>
<td>94</td>
<td>35</td>
<td></td>
</tr>
<tr>
<td>O395</td>
<td>O1, classical</td>
<td>Ogawa</td>
<td>1964</td>
<td>India</td>
<td>Patient</td>
<td>77</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>ZHE66</td>
<td>O1, classical</td>
<td>Ogawa</td>
<td>1963</td>
<td>China</td>
<td>Carrier</td>
<td>143</td>
<td>27</td>
<td></td>
</tr>
</tbody>
</table>
FIG. 1. ORFs absent from or present in the test strains compared with N16961. Final absent and present calls for each gene were translated to a binary code and analyzed with Cluster/TreeView. For each strain, a black line indicates the presence of a gene whereas a white line indicates its absence. The GC content curve is shown on the right.
<table>
<thead>
<tr>
<th>Strain</th>
<th>No. of genes in following category missing:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cell envelope</td>
</tr>
<tr>
<td>El Tor CTX⁺</td>
<td></td>
</tr>
<tr>
<td>V0518</td>
<td>1</td>
</tr>
<tr>
<td>SD95001</td>
<td>2</td>
</tr>
<tr>
<td>93097</td>
<td>1</td>
</tr>
<tr>
<td>V0550</td>
<td>11</td>
</tr>
<tr>
<td>19-22</td>
<td>11</td>
</tr>
<tr>
<td>JS32</td>
<td>8</td>
</tr>
<tr>
<td>7743</td>
<td>1</td>
</tr>
<tr>
<td>El Tor CTX⁻</td>
<td></td>
</tr>
<tr>
<td>D118</td>
<td>1</td>
</tr>
<tr>
<td>6312</td>
<td>0</td>
</tr>
<tr>
<td>79005</td>
<td>0</td>
</tr>
<tr>
<td>FJ80004</td>
<td>0</td>
</tr>
<tr>
<td>Wujiang-2</td>
<td>0</td>
</tr>
<tr>
<td>93284</td>
<td>0</td>
</tr>
<tr>
<td>V011506</td>
<td>0</td>
</tr>
<tr>
<td>200106</td>
<td>0</td>
</tr>
<tr>
<td>FJ147</td>
<td>1</td>
</tr>
<tr>
<td>ZJ0596</td>
<td>1</td>
</tr>
<tr>
<td>O139 CTX⁺</td>
<td></td>
</tr>
<tr>
<td>93010</td>
<td>6</td>
</tr>
<tr>
<td>93209</td>
<td>6</td>
</tr>
<tr>
<td>98106</td>
<td>6</td>
</tr>
<tr>
<td>FJ0299357</td>
<td>6</td>
</tr>
<tr>
<td>MO45</td>
<td>6</td>
</tr>
<tr>
<td>O139 CTX⁻</td>
<td></td>
</tr>
<tr>
<td>SCH04082</td>
<td>24</td>
</tr>
<tr>
<td>B4</td>
<td>25</td>
</tr>
<tr>
<td>94001</td>
<td>12</td>
</tr>
<tr>
<td>Classical</td>
<td></td>
</tr>
<tr>
<td>569B</td>
<td>1</td>
</tr>
<tr>
<td>O395</td>
<td>1</td>
</tr>
<tr>
<td>ZHE66</td>
<td>1</td>
</tr>
</tbody>
</table>
given strain results obtained with primers derived from flanking ORFs. The absence or presence of the ORFs was decided according to the PCR oligonucleotide probe ORFs with ratios between positive and negative cutoff differed from each other because of sequence divergence between strains. For identified according to its PCR result. The cutoff values in the various test strains between the positive and negative cutoff values was assayed by PCR and classified exactly to the highness and lowness of ratios. Finally, every ORF with a ratio greater than 0.6 were regarded as present. The presence or absence of an ORFs with ratios lower than 0.2 were classified as absent whereas those with a ratio between 0.2 and 0.6 were determined by PCR confirmation and mixed into 30 µl hybridization solution (3× SSC [1× SSC is 0.15 M NaCl plus 0.015 M sodium citrate], 0.2% SDS, 25% formamide, 5× Denhardt’s solution). DNA in the hybridization solution was denatured at 95°C for 3 min prior to loading onto a microarray. The microarray was hybridized at 42°C overnight and washed at room temperature with two consecutive washing solutions (0.2% SDS–2× SSC at 42°C for 5 min and 0.2% SSC for 5 min).

Imaging and data analysis. Microarrays were scanned with a LuxScan scanner (CapitalBio Corp.), and images were analyzed with SpotData software (CapitalBio Corp.). Analysis of CGH results was also accomplished by determination of fluorescence intensity ratios via a space- and intensity-dependent normalization procedure based on a LOWESS program. For each experiment, each ORF was spotted in triplicate onto each slide.

Classification of ORFs as absent or present in test strains. For each strain, ORFs with ratios lower than 0.2 were classified as absent whereas those with a ratio greater than 0.6 were regarded as present. The presence or absence of an ORF with ratios between 0.2 and 0.6 was determined by PCR confirmation assays to validate cutoff values as follows. About 80% of such ORFs in each strain were selected for a PCR assay. The positive cutoff was the highest ratio in all PCR-negative ORFs, while the negative cutoff was the lowest ratio of the PCR-positive ORF. Between the positive and negative cutoff values, some ORFs were PCR positive and others were negative; the PCR results did not correspond exactly to the highness and lowness of ratios. Finally, every ORF with a ratio between the positive and negative cutoff values was assayed by PCR and classified according to its PCR result. The cutoff values in the various test strains differed from each other because of sequence divergence between strains. For oligonucleotide probe ORFs with ratios between positive and negative cutoff values, the absence or presence of the ORFs was decided according to the PCR results obtained with primers derived from flanking ORFs.

Comparative phylogenomic analysis of microarray data. Final absent or present calls for each gene were translated into a binary code (gene present in a given strain = 2; absent = 0) and analyzed by Cluster/TreeView. Microarray data were processed with a practical extraction and reporting language procedure, and trees based on the unweighted-pair group method using average linkages (UPGMA) were constructed with BioNumerics (version 3.0; Applied Maths BVBA, Belgium). The correlations between these trees were calculated with BioNumerics.

FIG. 2. Variability of ORFs of different functional categories in different groups of V. cholerae strains. Each column in each category represents the ratio of the average number of absent genes within each V. cholerae group to the total number of genes in this category of N16961.

RESULTS

Overview of the array-based CGH results. Of the 3,773 ORFs surveyed, 2,876 were common to all strains, representing the functional core of the V. cholerae genome. However, 897 (23.8%) ORFs were variably present in the strains but the strains varied greatly from each other with respect to the number of missing ORFs. The majority of the absent ORFs were clustered together in regions of the N16961 genome that have a low GC content, suggesting possible gene transfer (Fig. 1).

The number of variant or missing genes in toxigenic El Tor strains is much lower (missing 9 to 34 genes) than in the other strains. Thus, the El Tor strains exhibit the highest genomic conservation and gene numbers, consistent with previous research (16), and the same conservation also exists in toxigenic O139 strains.

The large number of ORFs (112 to 599) missing from nontoxigenic strains is associated with a wide variety of diverse cellular functions. The variation in ORFs among the different V. cholerae strains was classified with respect to functional categories is shown in Table 2. Of the 2,876 common genes, the conservation categories are protein and amino acid biosynthesis (100 and 99.9%, respectively) in toxigenic El Tor strains (Fig. 2). Genes in the category of mobile and extrachromosomal element function exhibited high diversity, followed by those for cellular processes and central intermediary metabolism function, especially in nontoxigenic strains. In the mobile and extrachromosomal element category, subcategories that contain phage- and transposon-related genes appeared to be the most variable, suggesting a role in the diversification of V. cholerae. Of the 33 ORFs in the cellular process category that exhibit differences between toxigenic and nontoxigenic strains, 29 are pathogenesis-related genes.

The fluorescence intensity ratios of each gene of the toxigenic strains to N16961 were distributed around 1.0 and were clustered more narrowly than those of nontoxigenic strains.
Among nontoxigenic strains, the overall percentages of ratios of ORFs higher than 1.4 or lower than 0.4 in strains are greater than those of toxigenic strains, revealing that more sequence heterogeneity exists in nontoxigenic strains (Fig. 3).

A phylogenetic tree of the test strains was also generated on the basis of the microarray data (Fig. 4). All of the test strains could be divided into seven groups in accordance with serogroups, biotypes, TCP gene cluster possession, and tcpA alleles of El Tor derived or not, i.e., toxigenic El Tor strains, toxigenic O139 strains, classical strains, nontoxigenic TCP+ El Tor strains with an El Tor-type tcpA sequence (designated TCP\textsuperscript{ET}), nontoxigenic TCP+ El Tor strains with tcpA different from the El Tor type (designated TCP\textsuperscript{ET\textsubscript{var}}), nontoxigenic TCP− El Tor strains, and nontoxigenic O139 strains. In the group of TCP\textsuperscript{ET\textsubscript{var}} El Tor strains, the tcpA DNA sequences are different from those of the currently epidemic El Tor and O139 strains (GenBank accession no. AF512425, AF512423, and AY052831), and these strains were negative for the presence of the tcpA gene in microarray analyses.

Superintegron diversity and chromosome 2. The superintegron present in chromosome 2 of \textit{V. cholerae} is an active gene acquisition structure with frequent lateral gene transfers that are independent of the entire genome (25). Only five toxigenic El Tor strains, D118, 6312, 79005, FJ80004, and Wujiang-2, exhibited the same superintegron content as N16961, whereas all of the other strains showed diversity in this region.

Considering the specific characteristics of the superintegron and its diversity, we analyzed variation in the superintegron and its impact on chromosome 2, as well as on the whole \textit{V. cholerae} genome, by establishing phylogenetic trees with and without the superintegron. For this purpose, we constructed phylogenetic trees of the test strains based on the comparative hybridization data of the whole genome without the superintegron (I+II−SI) (Fig. 5B), of chromosome 2 alone (Fig. 5C), of chromosome 2 without the superintegron (II−SI) (Fig. 5A), and of the superintegron (SI) alone (Fig. 5D). The indexes of similarity between SI and I+II−SI and between SI and II−SI were found to be only 27.9 and 0%, respectively, indicating the instability of the superintegron and demonstrating its different evolutionary rate compared to the rest of genome. The similarity index also revealed frequent gene losses and gains in the superintegron, as well as recent evolution within this group of bacteria.

The index of similarity between the cluster analyses of chromosome 2 (II) and chromosome 1 (I) is 91.6%, and the index of similarity between chromosome 2 excluding the superintegron (II−SI) and chromosome 1 (I) is 93.7% (Fig. 5E). These results indicate that chromosome 2 without the superintegron has almost the same evolutionary rate as chromosome 1 if the superintegron region is excluded from consideration. The indexes of similarity between clusters (SI) and the whole genome and between SI and II are 45.5 and 59.3%, respectively. In the phylogenetic trees of the genomes without the superintegron, all of the toxigenic strains cluster in the same branching point,
similar to the tree based on sequences of multiple housekeeping genes (Fig. 6). This finding indicates that chromosome 2 without the superintegron is conserved and stable, as is chromosome 1.

**Comparison of serogroup O1 strains. (i) Toxigenic El Tor strains.** We observed conservation of genome content among the 11 toxigenic strains examined though they originated from diverse geographic locations and in different years (1961 to 2005). However, gene changes were observed among these strains according to the year of isolation. Thus, toxigenic El Tor from China can be divided into four subgroups that generally correspond to the times in which they were isolated (Fig. 4). The first subgroup contains a single strain, D118, isolated from a patient in 1961 when the seventh pandemic started. The only difference between D118 and other toxigenic El Tor strains was the absence of intact VSP-II in D118, a DNA region considered specific to seventh-pandemic strains. The second subgroup consists of five strains isolated between 1961 and 1980; within this group, strains 6312, 79005, Wujiang-2, and FJ80004 have the same genome content as N16961. The third subgroup consists of strains 93284, V011506, and 200106, which were isolated between 1993 and 2001; these strains are similar to N16961 but lack 16 ORFs. The fourth subgroup encompasses FJ147 and ZJ0596, which were isolated during the epidemic in 2005; 35 ORFs are absent from these two strains. Of the 35 absent ORFs, 18 (VC0495 to VC0512) are in VSP-II.

To further examine this phenomenon of missing ORFs, we sequenced the genome region corresponding to this missing fragment in FJ147 and ZJ0596. A new fragment was inserted between VC0495 and VC0513 (Fig. 7). BLAST analysis of this region revealed that it contained an IS element carrying a transposase with 100% identity to transposases in the N16961 genome; additionally, the flanking sequences suggest that this insertion was a random event.

(ii) Nontoxigenic El Tor strains. Nontoxigenic strains exhibited much higher diversity than toxigenic strains. The numbers of ORFs absent ranged from 112 (3.0% of the 3,773 ORFs detected in the array) for strain 7743 to 309 (8.2%) for strain JS32. These strains were divided into three clusters, (i) nontoxigenic TCPET strains (cluster I, strain 7743), (ii) nontoxigenic TCPETPV strains (cluster II, strains 93097, V0518, and SD95001), and (iii) nontoxigenic TCP strains (cluster III, strains JS32, 19-22, and V0550, which do not harbor VPI). These three clusters comprise different branches in the phylogenetic tree (Fig. 4). The number of ORFs absent from these clusters compared to N16961 increased from cluster I to cluster III (Table 2).

Cluster I includes one environmental strain, 7743, that has a VPI region with the same tcpA gene as toxigenic El Tor strains. In several biochemical tests, including sensitivity to the V. cholerae typing phages, sorbitol and mannitol fermentation, and hemolysis activity, strain 7743 shows the same characteristics as toxigenic El Tor strains but differs from the nontoxigenic strains (data not shown). A total of 112 ORFs were absent from strain 7743. In comparison, on average, only 12 ORFs were absent from toxigenic El Tor strains and 273 ORFs were absent from cluster II and III nontoxigenic El Tor strains. Moreover, in the phylogenetic tree, 7743 is clustered together with toxigenic strains rather than with other nontoxigenic ones, and the cluster III strains diverge strongly from the epidemic strains (Fig. 4).

Aside from VPI, we also found sequence variations between cluster II and cluster III nontoxigenic strains in other genome regions. One gene cluster, VC0919 to VC0925, contains exopolysaccharide (EPS) biosynthesis genes. This region was
FIG. 5. Phylogenetic structures (UPGMA trees) of the strains according to microarray data of (A) chromosome 2 excluding the superintegron (II–SI), (B) the whole genome excluding the superintegron (I+II–SI), (C) chromosome 2 (II), (D) the superintegron (SI), and (E) chromosome 1 (I). The indexes of similarity between different phylogenetic trees were calculated by Dice tests with BioNumerics and are listed in panel F. The strain codes represent toxigenic El Tor strains, group A, including 6312 (A1), 79005 (A2), FJ80004 (A3), N16961 (A4), D118 (A5), Wujiang-2 (A6), 93284 (A7), VL011506 (A8), 200106 (A9), FJ147 (A10), and ZJ0596 (A11); toxigenic O139 strains, group B, including MO45 (B1), 93010 (B2), 93209 (B3), 98106 (B4), FJ0299357 (B5), and 200306 (B6); classical strains, group C, including 569B (C1), O395 (C2), and ZHE66 (C3); nontoxigenic TCPET El Tor strains, group D, including 7743 (D1); nontoxigenic TCPETvar strains, group E, including 93097 (E1), SD95001 (E2), and V0518 (E3); nontoxigenic TCPET El Tor strains, group F, including 19-22 (F1), JS32 (F2), and V0550 (F3); and nontoxigenic O139 strains, group G, including 94001 (G1), B4 (G2), and SCH04082 (G3).
absent from cluster III nontoxigenic El Tor strains but present in cluster II nontoxigenic El Tor strains by hybridization. However, amplicons with the same size as that of N16961 in cluster III nontoxigenic El Tor strains were obtained by using primers for the flanking ORFs of VC0919 to VC0925. Restriction enzyme fragment lengths of PCR products of this region revealed different patterns for cluster III nontoxigenic El Tor strains and N16961. Additionally, cluster III strains also varied among themselves. These data indicate that multiple new sequence alleles exist in this region. Since the EPS are involved in biofilm formation and motility, as well as in the secretion of hemolysin and CT, in V. cholerae (34), variance of these genes might suggest differences in EPS roles and environmental survival among nontoxigenic TCP⁻, TCP⁺, and toxigenic El Tor strains.

Several other virulence-related genes, including ompU and pilA, also exhibited sequence variation within cluster I nontoxigenic strains compared to seventh-pandemic strain N16961. In the array analysis, these genes in cluster III strains were shown to be absent whereas adjacent genes were present. We sequenced the corresponding regions and found that the ompU sequences in cluster III strains diverged at the nucleotide level. Thus, the sequence identity was 66.4% (strains 19-22 and V0550, 73.7% amino acid sequence identity) and 74.1% (strain JS32, 79.5% amino acid sequence identity) with respect to N16961. OmpU is a pore-forming protein of the outer membrane of V. cholerae. Since V. cholerae lacking ompU has reduced tolerance to bile salts, it is possible that strains with variant ompU alleles differ in environmental survival ability (33). Similarly, pilA of cluster III strains was only 59.8% identical, at the nucleotide level (11 to 49% amino acid sequence identity only), to pilA of N16961. Because pilA contributes to the adherence and colonization abilities of bacteria (51, 18), such allelic variation may have significance for colonization. The DNA sequences of ompU and pilA of the cluster II strains and strain 7743 were identical to the sequence of N16961.

Another region where sequence variation was observed between cluster II and III nontoxigenic El Tor strains was the superintegron. Other differences were found. For instance, part of VPI-II (VC1773 to VC1788) is present in cluster III of nontoxigenic El Tor strains but absent from cluster II strains. Likewise, fragments of VC1748 to VC1753 and VCA1042 to VCA1044 are present in cluster II strains but absent from cluster III strains. VC1748 to VC1753 encode pqiA and five hypothetical proteins, while VCA1042 to VCA1044 encode the Ccm2-related protein and the TagE protein.

On the basis of the array constructed with the genome content of N16961, in all tested toxigenic and nontoxigenic El Tor strains, a total of 455 ORFs were missing from various El Tor strains. Of these, 31 are commonly found only in toxigenic El Tor strains. Thirty of these common genes are genomic island
components, including VSP-I, CTXΦ, and the superintegron. The remaining ORF, VC2346, is annotated as encoding the Smp protein, but its function is unknown.

Comparison of genome contents of serogroup O139 strains.

(i) Differences between the genome contents of toxigenic El Tor and O139 strains. It was likely that the serogroup O139 V. cholerae strains originated from an El Tor-like strain (53), and the results of our whole-genome content analysis support this idea (Fig. 4). Compared to toxigenic El Tor strains, 77 ORFs were absent or highly divergent in all toxigenic O139 strains and most of them are clustered in three genomic islands and in the gene cluster encoding the O antigen. The first island represents 83% of the gene cluster encoding enzymes involved in lipopolysaccharide antigen synthesis (Fig. 8). The second island spans the region from VC1761 to VC1787 and consists of two parts of VPI-II, including an ORF that encodes neuraminidase. Neuraminidase may increase host cell susceptibility to cholera toxin (23). Possible explanations for the reemergence of El Tor cholera in Bangladesh since 1994 (19) are the presence of VPI-II in the El Tor biotype and the deletion of most of VPI-II from O139 strains (30). The third island is the superintegron in which 19 ORFs are missing from toxigenic O139 strains. Herein we show that the genes missing from the superintegron of toxigenic O139 strains are almost the same as those of toxigenic El Tor strains isolated after 1993 (93284, 200106, V011506, ZJ0596, and FJ147). Moreover, the first three toxigenic El Tor strains are clustered more closely with toxigenic O139 strains in the phylogenetic tree based on the superintegron hybridization profile (Fig. 5D), suggesting a close relationship between toxigenic O139 and El Tor strains prevailing at that time.

(ii) Divergence of nontoxigenic O139 strains. In contrast to the conservation of toxigenic O139 strains, the nontoxigenic O139 strains showed more diversity in both the number and categories of absent ORFs (Table 2). These absent ORFs are dispersed on the two chromosomes. The toxigenic and nontoxigenic O139 strains have the same microarray hybridization patterns in the gene cluster involved in lipopolysaccharide antigen synthesis, confirming the genetic basis of the serogroup antigen of these strains (Fig. 8). Nonetheless, in the phylogenetic tree, the nontoxigenic O139 strains are scattered throughout the tree, indicative of the large genetic diversity of genomic content among these strains.

DISCUSSION

In this study, array-based CGH was used to investigate the genome diversity of serogroup O1 and O139 V. cholerae, including toxigenic and nontoxigenic strains. In the comparison, we used seven nontoxigenic O1 and three nontoxigenic O139 strains. Analysis of the difference in genome components of nontoxigenic strains compared to toxigenic strains may facilitate tracing of the origin and evolution of seventh cholera pandemic strains.

Conservation of chromosome 2 and impact of the superintegron on strain diversification. It is unclear why V. cholerae has two chromosomes. Sequence comparison of chromosomes 1 and 2 (strain N16961) indicates that chromosome 2 was originally a megaplasmid captured by an ancestral Vibrionaceae species; perhaps chromosome 2 provides certain evolutionarily selective advantages for the bacteria (25). Our examination of multiple strains (including nontoxigenic O1 and O139 strains) indicates that the proportion of absent genes in chromosome 2 is higher than is the case in chromosome 1. Our analysis also shows that the majority of missing genes are located within the superintegron, a region of more than 200 ORFs. Phylogenetic trees of these strains, based on chromosome 2 but excluding the superintegron, reveal that the gene content of chromosome 2 is as conservative as that of chromosome 1. Therefore, our CGH data suggest that chromosome 2 (aside from the superintegron) is conservative in V. cholerae. The presumed reason that the phylogenetic trees for chromosomes 1 and 2 do not correspond strictly is that the superintegron provides variability in genome content.

Phylogenetic trees based on microarray hybridization results of chromosome 2 and the superintegron alone revealed evolutionary trends (shared or otherwise) between toxigenic O1 and O139 strains and between toxigenic O1 El Tor strains. The frequent gene gains and losses in the superintegron, characteristic of such integrons, represent recent genetic change and active gene transfer. It has been reported that microevolution is an ancient and perpetual event in Vibrionaceae resulting in the loss, gain, and duplication of genes (47), and our results generally support that observation. The structural diversification of the genomes of toxigenic El Tor strains spanning 45 years in this study favors this viewpoint.

The superintegron cassettes examined to date seemingly encode adaptive rather than indispensable functions (48). In the present study, the nontoxigenic strains exhibited greater diversity than the toxigenic strains in the superintegron region. The impact of integrons on bacterial evolution remains to be eval-
uated beyond their obvious role in the dissemination of antibiotic resistance genes. The variation of the superintegron and its impact on evolution of *V. cholerae* can be explained more accurately only after elucidation of the function of the genes inside the structure.

**Microevolution among the seventh-pandemic toxigenic El Tor strains.** The seventh cholera pandemic was caused by toxigenic El Tor strains, which were reported to be a conserved clone (16). In spite of the observation noted in the 11 toxigenic El Tor strains, minor differences among these strains can be observed and have accumulated over time from 1961 to now. In the phylogenetic tree for chromosome 2 (Fig. 5), toxigenic strains can be clustered into different groups according to the year of isolation, reflecting ongoing but exiguous evolution.

VSP-I and VSP-II are the characteristic genome islands identified in the seventh-pandemic El Tor strains and toxigenic O139 strains (16, 44, 45). In this study, a toxigenic strain, D118, was found to lack VSP-II entirely. It is reported that two toxigenic El Tor strains isolated in the 1970s lacked both VSP-I and VSP-II or only VSP-II, and one toxigenic El Tor strain from 1937 lacked VSP-I, VSP-II, and VPI-2 (45). Such strains resemble the presumed progenitors of the seventh-pandemic El Tor strains (16, 45). Strain D118, used in this study, was isolated in 1961, the year the seventh pandemic began. In the phylogenetic tree based on the sequences of four housekeeping genes, D118 is in a lineage separate from that of other toxigenic El Tor strains (Fig. 6). Therefore, D118 may represent a strain or group of strains that were converted into the predominant clone by the acquisition of VSP-II.

However, two epidemic strains isolated in 2005, FJ147 and ZJ0596, were found to lack most of the ORFs in VSP-II, and these two strains caused epidermias spanning thousands of kilometers and lasting several months in southeastern China. Therefore, the function of VSP-II in bacterial pathogenesis, environmental survival, and epidemic instigation remained to be defined. Likely the seventh-pandemic strains are in continual evolution, and it is expected that they may evolve into new pathogenic clones through the acquisition of new virulence and/or adaptive factors.

**Diversity of nontoxigenic strains and clonality of toxigenic strains in serogroups O1 and O139.** Previous studies revealed high diversity in nontoxigenic strains, especially the non-O1 and non-O139 isolates, through molecular subtyping techniques including ribotyping, pulsed-field gel electrophoresis, housekeeping gene sequencing, and amplified fragment length polymorphism (20, 32, 45, 50). Herein we used microarray methods to investigate the genome contents of toxigenic and nontoxigenic O1 and O139 strains. Extensive genomic diversity among these strains was again found.

Comparative analyses of nearly 300 microbial species have demonstrated that multiple forces have shaped prokaryotic genomes during their dynamic evolution (22). The genomic diversity of toxigenic and nontoxigenic O1 and O139 *V. cholerae* strains reflects the same dynamism of these other prokaryote genomes.

Toxigenic O1 and O139 *V. cholerae* strains exhibit greater clonality than environmental non-O1/non-O139 strains (32, 45). Our study supports this observation by demonstrating that the genomic content of nontoxigenic *V. cholerae* strains is quite different from that of closely related toxigenic strains. It has been suggested that populations of bacteria typically consist of abundant and heterogeneous isolates that rarely cause disease but that a small number of groups of closely related strains (clones or lineages) are particularly associated with epidemics of disease (40). Serogroup O37 caused an outbreak in 1968 (58). IS1004 fingerprinting, housekeeping gene sequencing, and multilocus enzyme electrophoresis analysis revealed that epidemic O37 strains are phylogenetically clustered with O1 and O139 isolates and that they are especially close to O1 classical strains (3, 6, 45). It seems that, to date, cholera epidemics have been caused only by strains possessing an O1 serogroup core genome, despite the fact that O37 and O139 are non-O1 serogroups (36, 45).

A study in an area where cholera is endemic provided a model for understanding the selection of toxigenic strains from the genetically diverse nonpathogenic strains in the environment; such selection has been suggested to occur through enrichment in the mammalian host intestine (20). In areas where cholera is endemic, large numbers of toxigenic *V. cholerae* cells are present in the stool of patients which, when excreted into environmental waters, contribute in part to the expansion of cholera epidemics. A study indicated that by passing through the mammalian intestine, *V. cholerae* cells acquire a hyperinfectious state that may also contribute to the epidemic spread of cholera (41). We suppose that the rapid spread of toxigenic *V. cholerae* strains in areas where cholera is endemic makes the isolates in epidemics more conservative in clonality. In contrast, nontoxigenic strains of diverse lineages do not undergo such enrichment and exhibit extensive genetic diversity. It should also be considered whether the epidemic strains have advantages in maintenance and multiplication over nonpathogenic strains in aquatic environments.

Our CGH analysis revealed that the O139 and El Tor strains isolated in the coepidemic time period (10 more years since the emergence of O139 *V. cholerae*) exhibit superintegron missing-gene profiles essentially identical to those of El Tor strains isolated before the emergence of O139 strains. It has been proposed that serogroup O139 originated from an El Tor strain that obtained the O139 antigen biosynthesis and some other gene clusters (4, 52, 55). However, some portion of VPI-II is absent from all of the test toxigenic O139 strains isolated through the past years, and an additional five common missing genes in the superintegron in the toxigenic O139 strains have been documented (in comparison with all toxigenic El Tor strains). We speculate that these missing genes were lost after the replacement of the O-antigen-synthetic gene cluster.

Genomic analysis of toxigenic and environmental nontoxigenic O1, O139, and non-O1/non-O139 *V. cholerae* strains can provide clues to *V. cholerae* evolution and the origin of pathogenic strains. Strains that can trigger epidemics may acquire virulence genes successively (9, 10, 20, 45), rather than in a single horizontal gene transfer event. In this study, we classified the nontoxigenic O1 El Tor *V. cholerae* strains into three clusters, named TCP⁺, TCP⁻, and TCP⁻Ent strains. We speculate that the nontoxigenic TCP⁻Ent and TCP⁻Entvar strains represent intermediate strains in different lineages evolving from nontoxigenic to toxigenic strains. The acquisition of ctxAB through the lysogenic conversion of CTXφ may confer the ability to initiate epidemics upon those strains (54). The gene
of pIII\textsuperscript{CTX} (previously known as OrfU) of CTX\Phi encodes a protein that can recognize TcpA as the receptor of phage infection (26); however, tcpA has different sequence variants (12, 24, 28, 42, 43, 46). Variable gene sequences of pIII\textsuperscript{CTX} have also been documented (9, 11). The tcpA sequence of nontoxicogenic strain 7743 (also known as IEM101) used in this study is identical to the sequence of seventh-pandemic El Tor strains, and the attB site for CTX\Phi genome integration is also identical (57). Strain 7743 can be infected by CTX\textsuperscript{ET}\Phi and thereby converted into a toxigenic strain (37, 38, 57). Moreover, rabbits immunized with 7743 and the live-vaccine candidates developed from 7743 can be protected against a challenge with classical or El Tor strains (37–39, 57). This protection demonstrates that strain 7743 has the same immunogenicity as toxigenic El Tor strains. Intact VPI, VPI-II, and TCL clusters are present in strain 7743, whereas VSP-I, VSP-II, and CTX\Phi are absent. Taken together, these data suggest that it is possible that strain 7743 represents an intermediate in the origination of the seventh-pandemic clone.

It has been reported that the different TcpA alleles can serve as CTX\textsuperscript{ET}\Phi and CTX\textsuperscript{ak}\Phi receptors. However, a difference in the infection frequency of CTX\Phi among these strains has been documented (11); perhaps there are barriers to the cross-infection with some CTX\Phi with different pIII\textsuperscript{CTX} alleles and nontoxicogenic strains with different tcpA alleles. Perhaps, in comparison with strain 7743, nontoxicogenic TCP\textsuperscript{ETvar} strains may have a reduced ability to acquire CTX\textsuperscript{ET}\Phi. Such strains may provide mechanistic clues about evolutionary events in the development of toxigenic strains. In other words, perhaps these strains (or similar strains) have the potential to become new dominant clones in future epidemics through the acquisition of CTX\Phi with the corresponding pIII\textsuperscript{CTX} allele.

Obvious divergence in virulence-related genes such as tcpA (12, 21, 24, 28, 42, 43, 46), the gene for pIII\textsuperscript{CTX} (9, 11), and rstR (9, 14, 15) in CTX\Phi has been found in different strains. Similarly, sequence variance in the ompU, pilA, and EPS cluster was also observed in nontoxicogenic O1 \textit{V. cholerae} in our study. It is inconvincible that this variation represents an accumulation of random mutations because the ORFs flanking those genes show little or no nucleotide variation.

\textit{V. cholerae} can survive in aquatic environments and host intestines. Different gene expression mechanisms are used (7, 27, 35, 41, 56) in these two niches. Besides the modulation of gene expression, in different environments, gene sequence variation may confer selective advantages on the strains in various environments. The hypervariability at the tcpA locus is thought to be generated through multiple horizontal transfer and recombination events, and tcpA diversity is presumed to reflect diversifying selection resulting from the host immune response, susceptibility to CTX\Phi (11). The sequence variability in type IV pili of several other bacterial species has also been reported (8, 11). Therefore, we speculate that variation in the ompU, pilA, and EPS gene cluster may contribute to functional differences (phenotypes) in \textit{V. cholerae} as the strains respond to selective pressure.

**Conclusion.** Our analysis shows that nontoxicogenic strains from serogroups O1 and O139 are a collection of heterogeneous clones that are part of the natural microflora of aquatic environments. Some strains carry virulence-related genes that differ in sequence from those of the seventh-pandemic El Tor strains; although these strains have no El Tor-type TCP, they may nonetheless have suitable receptors that allow infection by other types of CTX\Phi. For example, the CTX\Phi with pIII\textsuperscript{CTX} differed from CTX\textsuperscript{ET}\Phi and may infect the nontoxicogenic strains with tcpA sequences distinct from that of the El Tor type. The possibility is proposed that these strains can serve as raw material for the development of (future) epidemic clones through acquisition of novel alleles of CTX\Phi. These nontoxicogenic O1 and O139 strains may serve as models to understand how environmental nontoxicogenic strains become toxigenic strains via gene loss and gain. Array-based CGH provides the ability to rapidly examine the genomic contents of numerous bacterial isolates, thereby allowing determination of which genes are present and which are absent in nontoxicogenic strains compared to toxigenic ones, a powerful tool. We proposed that more robust conclusions regarding the pathogenicity, environmental adaptation, and evolution of \textit{V. cholerae} can be achieved by whole-genome sequencing and biological research on these nontoxicogenic strains.

**ACKNOWLEDGMENTS**

This research was supported by the National Basic Research Priorities Program (grant G1999054102 from Ministry of Science and Technology) and the Key Technologies R&D Program (grant 2003BA712A05-04 from the Ministry of Science and Technology, People’s Republic of China).

**REFERENCES**

17. Dziejman, M., E. Balon, D. Boyd, C. M. Fraser, J. F. Heidelberg, and J. J.


