Sequence Variation in tcdA and tcdB of Clostridium difficile: ST37 with Truncated tcdA Is a Potential Epidemic Strain in China

Pengcheng Du, Bo Cao, Jing Wang, Wenge Li, Hongbing Jia, Wen Zhang, Jinxing Lu, Zhongjie Li, Hongjie Yu, Chen Chen, Ying Cheng

Key Laboratory of Surveillance and Early Warning on Infectious Disease, Division of Infectious Disease, Chinese Center for Disease Control and Prevention, Beijing, China; State Key Laboratory for Infectious Disease Prevention and Control, and National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China; Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Hangzhou, China; Department of Clinical Laboratory, China-Japan Friendship Hospital, Beijing, China

Clostridium difficile is a well-known nosocomial infectious pathogen. Research on C. difficile infection has primarily focused on strains such as the hypervirulent PCR ribotype 027 (sequence type 1 [ST1]) emerging in Europe and North America. However, other new emerging ribotypes in some countries have attracted attention, such as PCR ribotype 17 (ST37) in Asia and Latin America. We collected 70 strains and sequenced their toxin genes, tcdA and tcdB. Multi locus sequence typing (MLST) was used to study their population structure. In addition, tcdA and/or tcdB sequences of 25 other isolates were obtained from GenBank. Single nucleotide polymorphisms (SNPs) were identified and analyzed. Phylogenetic analyses were performed to study toxin gene evolution. All tcdA and tcdB sequences were divided into 1 of 16 types (denoted A01 to -16 and B01 to -16, respectively).

Hypervirulent strain RT027 is A13B12, and RT078 is A14B10, whereas the newly epidemic strain RT017 is A15B13. SNP analysis suggests the possibility of recombination in tcdB, perhaps through horizontal gene transfer. SNPs were also found in the sequences corresponding to the PCR primers widely used for toxin detection. Our study shows that ST037 shares a few genotypic features in its tcdA and tcdB genes with some known hypervirulent strains, indicating that they fall into a unique clade. Our findings can be used to map the relationships among C. difficile strains more finely than can be done with less sensitive methods, such as toxityping or even MLST, to reveal their inherent epidemiological characteristics.

C. difficile is a nosocomial bacterial pathogen that causes antibiotic-associated diarrhea mediated by cellular exotoxins secreted into the intestine during bacterial growth (1,2). In the past decade, the mortality from C. difficile infection (CDI) has increased from 6% to 13.5% overall among older patients (3). Research on CDI has primarily focused on strains such as the hypervirulent PCR ribotype 027, multilocus sequence type 1 (RT027, ST1), emerging in Europe and North America, which produces two major toxins, A and B, encoded by the genes tcdA and tcdB in the pathogenicity locus (PaLoc) (4,5). In recent years, however, the epidemiology of CDI has changed dramatically, since other emerging PCR ribotypes have become prevalent (6-13), and several pathogenic A-negative B-positive (A - B+) strains (which produce toxin B, but not toxin A) have appeared in Asia and Latin America (8-12, 14, 15). Hence, it has become a matter of urgency to understand the new, complex pattern of A - B+ C. difficile strain variants and to reveal their relationships to other epidemic genotypes.

Recently, a particular variant strain of C. difficile, ST37, which has been ribotyped as RT017 and produces toxin B only, has attracted increasing attention (11, 12, 16). This type of strain shows increased resistance to clindamycin and erythromycin, with concomitant greater risk to inpatient health (13, 17, 18). Currently, the literature on CDI caused by A - B+ strains is limited, and there is some confusion over the correct identification of specific isolates. For example, the apparently high isolation rate of A - B+ strains in Asia and Latin America may reflect mismatching of PCR primers as a result of C. difficile polymorphisms (19). Nevertheless, it seems clear that C. difficile RT017 is prevalent in Asia and China, but the reasons for this distribution, as well as its relation to the highly virulent RT027 strain in Europe and North America, are still uncertain.

To address this situation, we collected C. difficile strains from five different geographical areas in China. The entire tcdA and tcdB genomic regions were sequenced and single nucleotide polymorphism (SNP) analysis was conducted (i) to provide information on genomic diversity and to aid PCR primer design for later epidemiological studies, (ii) to reveal the potential linkage between RT027 and RT017, and (iii) to throw new light on the widespread occurrence of RT017 in China. Our data also provide reference data for tcdA and tcdB sequences from the sampled regions, which will enable transmission patterns to be determined and allow forecasting of future distribution trends in these areas of China.

MATERIALS AND METHODS

Strain isolation and DNA extraction. Sixty-six of the 70 C. difficile strains analyzed derived from five provinces and municipalities of China, including Beijing (21), Shanghai (15), Shandong (11), Henan (7), and Guang-
The strains in each province were collected from one general hospital and randomly selected regardless of the years of isolation to include as many strains as possible. The remaining four strains were isolated from the United States, the United Kingdom, Japan, and France (see Table S1 in the supplemental material). Stool specimens from diarrheal patients were collected using Transwabs (MW&E Ltd., Wiltshire, England), and then cultured on cycloserine-cefoxitin-fructose agar (CCFA) with 5% egg yolk. Colonies that demonstrated a typical morphology (flat, yellow, ground-glass appearance) and odor on the CCFA and in Gram staining were Gram-positive bacilli with subterrestrial spores, and those that yielded positive results in response to the commercially available latex agglutination test (Oxoid, Ltd., Basingstoke, United Kingdom) were identified as \textit{C. difficile}. Isolates that were not confirmed by these methods were further identified via API 20A (Bio Mérieux, France) and 16S rRNA gene sequencing (with the primer set 5'-GGAGGACGCAGTGGGGATGAAT A-3' [forward] and 5'-TGACGGGCCGCTTGTGTAACAGG-3' [reverse]) and glucose dehydrogenase (GDH) gene amplification and sequencing (6) (with the primer set 5'-TTCCTAATTTAGCAGCAGCTTC-3' [forward] and 5'-GTCTTGGATGGTTGATGAGTAC-3' [reverse]).

Characterization of toxin genes by PCR and multilocus sequence typing. The toxigenic property of each \textit{C. difficile} isolate was determined by characterization of the \textit{tcdA} and \textit{tcdB} genes. A PCR assay for \textit{tcdB} was performed using primers NK104 and NK105, which resulted in a 203-bp amplicon for \textit{tcdA}-positive strains (9). The \textit{tcdA} gene was detected using primers \textit{tcdA-F} and \textit{tcdA-R}, which yielded a 369-bp amplicon for \textit{tcdA}-positive strains and a 110-bp amplicon for \textit{tcdA}-negative strains (20). All 70 strains were characterized by the MLST method with seven housekeeping genes (21), and sequences were submitted to the \textit{Clostridium difficile} multilocus sequence typing (MLST) database (http://pubmlst.org /edifice) to acquire a sequence type (ST). Thirty-five strains were part of the 104 strains used in our previous MLST study (16). A minimum spanning tree was also constructed to exhibit the population structure of Chinese strains using the categorical data for MLST via BioNumerics v4.0 software (Applied Maths BVBA, Belgium).

Sequencing of \textit{tcdA} and \textit{tcdB}. Using a primer-walking method, nine pairs of primers for \textit{tcdA} and eight pairs for \textit{tcdB} were designed via Primer Premier 5.0 software to cover the whole length of each gene (8,133 bp and 7,101 bp, respectively) in overlapping segments. The sequence of \textit{tcdA} and \textit{tcdB} was a Chinese \textit{A} \textit{B} strain with numerous \textit{tcdA} SNPs distinguishing it from other Chinese \textit{A} \textit{B} strains. Strains with \textit{A13} (AG3) and \textit{A14} (AG4) existed only in Europe and North America. A detailed analysis of functional domains encoded by \textit{tcdA} revealed that mutations mainly occur in the receptor-binding domain, and largely consist of nonsynonymous SNPs. The numbers and rates of nonsynonymous SNPs in the receptor-binding domain are significantly higher than in the other three domains, indicating that this region is undergoing rapid evolution.

Two variant gene types (A15 and A16), with truncated or deleted sequences and mutation rates near 2.0% in \textit{tcdA}, were found in \textit{A} \textit{B} strains and assigned to groups AG5 and AG6, respectively. A16 was an unusual sequence from strain 8864 that was only 2,091 bp long. Nonsense mutations were found in both groups, but compared to A16, the premature stop codon in A15 appeared much earlier, along with an approximately 6-kb nonfunctional region containing a large number of mutations. All 22 ST37 strains contained A15-type \textit{tcdA} genes.

\textit{tcdB} sequence analysis. We obtained 16 \textit{tcdB} variants from 95 sequences, of which five sequence types (B04, B05, B07, B08, and B09) were new compared to those in GenBank. B09 was derived from the same Chinese \textit{A} \textit{B} strain that contained the highly mutated \textit{tcdA} type A12 and also contained multiple \textit{tcdB} sequences in \textit{tcdB} (Fig. 2; also see Table S4 in the supplemental material). There were

\textbf{RESULTS}

Overview of representative sequenced strains. Sixty-six Chinese strains used in this study were isolated from the 1980s to 2012, and 42 of them (63.6%) were from inpatients. All 70 individual \textit{C. difficile} isolates include the sequence types 1 (1 strain) 2 (11 strains), 3 (12 strains), 8 (2 strains), 35 (13 strains), 37 (18 strains), 46 (1 strain), 53 (1 strain), 54 (4 strains), 55 (2 strains), 92 (1 strain), 99 (1 strain), 102 (1 strain), 129 (1 strain) and 221 (1 strain). Their \textit{tcdA} and \textit{tcdB} genes were successfully amplified and sequenced. In total, 24 \textit{tcdA} and \textit{tcdB} sequences of 25 further \textit{C. difficile} isolates were distilled from GenBank (see Table S1 in the supplemental material), for 10 of which only one of the two genes was available and in other cases the ST types were uncertain. In total, 25 \textbf{A} \textbf{B}*, 63 \textbf{A} \textbf{B}*, and 7 strains of unknown toxin status were analyzed. All strains were divided into 16 STs by MLST analysis, of which ST11 was the only type not found in China in this study (see Table S1 in the supplemental material). Unlike in Europe and the United States, in China the predominant type in our sampled strains is ST37 (which is quite different from other STs) followed by ST35. All \textbf{A} \textbf{B}* strains obtained in our study are ST37, and \textit{tcdA} from these strains contains partially truncated sequences.

\textit{tcdA} sequence analysis. In total, 16 \textit{tcdA} variants were obtained from 85 sequences, of which 10 sequence types (A02, A03, A04, A05, A06, A07, A08, A09, A11, and A12) from Chinese strains were new compared to those in GenBank; multiple SNPs were found only in A12 (Fig. 1; also see Table S3 in the supplemental material). There were 14 different sequence types (A01 to A14) for \textit{tcdA} from \textbf{A} \textbf{B} strains, which clustered into four groups (AG1 to AG4). AG1 is the largest group. It contained 11 sequence types in which the average number of SNPs was 9.3, with mutation rates all below 0.2% (Fig. 1). In the other three \textbf{A} \textbf{B} groups (AG2 to AG4), a large number of SNPs distinguished the respective strains, which had mutation rates of about 1.5%. The \textit{C. difficile} isolate containing sequence type A12 (AG2) is a Chinese \textbf{A} \textbf{B} strain with numerous \textit{tcdA} SNPs distinguishing it from other Chinese \textbf{A} \textbf{B} strains. Strains with \textit{A13} (AG3) and \textit{A14} (AG4) existed only in Europe and North America.
There are 12 different tcdB sequence types (B01 to B12) in the A⁺ B⁺ strains, which are clustered into four groups (BG1 to BG4). BG1 was the largest group. It contained eight sequence types (B01 to B08), in which there were on average 2.4 mutations per sequence; all mutation rates were below 0.1%. The remaining four sequence types (B09 to B12) were divided into the other three A⁺ B⁺ groups (BG2 to BG4) and showed mutation rates of about 1.35%, 2.85%, and 6.55%, respectively (Fig. 2). While AG2 to AG4 shared a number of mutations with each other, BG2 to BG4, which comprised many of the same strains found in AG2 to AG4, had far more diversity in their SNPs. BG2 possessed fewer mutations than BG3, which in turn had far fewer SNPs than BG4. Most SNPs, the majority of which in these groups were nonsynonymous, were concentrated in the delivery domain and receptor-binding domain.

Four tcdB sequence types (B13 to B16) found in the A⁻ B⁺ strains were 7,104 bp in length, which differed from the 7,101 bp seen in most A⁻ B⁺ strains. B13, B14, and B15 clustered into the same group (BG5), with an average mutation rate above 5.0% and with SNPs mainly distributed in the first two domains, which have glucosyltransferase and autoprotease functions. All 22 ST37 strains contained the B13 tcdB sequence type. B14 and B15 differed only slightly from B13.

Interestingly, BG6, comprising sequence type B16 only, contained SNPs throughout its length and had the highest mutation rate (11.38%). In its first two domains, BG6 possessed mutations that were similar to those in BG5. However, BG6 mutations in the third and fourth domains, which have delivery and receptor-binding functions, resembled mutations found in BG4, particularly in the fourth domain. Although the first two domains in BG5 (A⁻ B⁺ [RT017]) were very similar to those of BG6 (A⁻ B⁺ [8864]), its last two domains resembled those of BG1 (A⁻ B⁺ [common]). In another example of this phenomenon, the last 3 kb of the tcdB variant defining BG4 (A13B12 [RT027]) also showed a high degree of identity to the corresponding region of BG6 (A⁻ B⁺ [8864]), but the first 4 kb of BG4 was more similar to that of BG3 (A⁻ B⁺ [RT078]), which is 203 SNPs (95% similarity) with BG3 versus 420 SNPs (90% similarity) with BG6. Taken together, this analysis suggests the possibility of recombination in tcdB, perhaps through horizontal gene transfer (HGT) of the toxin genes or the whole PaLoc. This may be significant in the context of the evolution of hypervirulent strains.

Molecular characteristics of tcdA and tcdB sequence types in epidemic strains. Based on sequence similarity and phylogenetic analysis, the 16 tcdA types and 16 tcdB types were clustered into six
clades, in keeping with the six groups of both \textit{tcdA} and \textit{tcdB} variants (Fig. 3). The epidemic strains can be analyzed with respect to their toxin gene sequence type. Thus, the hypervirulent strain RT027 (ST1), which presents as A$^+$/B$^+$/H11001 and is binary toxin positive, is included in clade 2, with toxin gene sequence type A13B12.

Epidemic strain RT078 (ST11), which is also A$^+$/B$^+$/H11001 and binary toxin positive, is a member of clade 4, with toxin gene sequence type A14B10. An examination of the phylogenetic relationship between RT027 and RT078, based on their \textit{tcdA} and \textit{tcdB} sequences, showed that the toxin genes in these strains are evolving separately.

Epidemic strain RT017 (ST37), which presents as A$^-$ B$^+$ and has a high recent isolation rate in Asia and Latin America, forms part of clade 3, with toxin gene sequence type A15B13. Multiple nonsynonymous mutations downstream of the glucosyltransferase domain were found in its \textit{tcdB} gene, together with a nonsynonymous mutation at nucleotide position 139 (i.e., C139T) in its \textit{tcdA} gene, generating a premature stop codon in the TcdA amino acid sequence (Q47X). Currently, the major strains in China are ST37, but ST37 isolates have also been recorded in other countries, including Ireland, Belgium, and the United States. Furthermore, the sequences of both \textit{tcdA} and \textit{tcdB} from all 22 ST37 strains isolated worldwide since the 1980s are identical, suggesting that ST37 may be a rapidly spreading sequence type. Whole-genome analysis also showed a high degree of similarity between ST37 strains (24, 25).

**DISCUSSION**

Toxin A and B mediate the main clinical symptoms of CDI, consistent with their coding genes within the same genomic region, and are expressed together (26, 27). However, the newly prevalent cases caused by A$^-$ B$^+$ strains appear to challenge the requirement for both toxins in the etiology of this disease. According to the size of deletions at the 3' end of \textit{tcdA}, A$^-$ B$^+$ strains are divided into

---

**FIG 2 Identification of single nucleotide polymorphisms (SNPs) in \textit{tcdB} sequences using VPI 10463 as reference. The coding regions of the gene are shown as four domains, glucosyltransferase, autoprotease, delivery domain, and receptor-binding domain, according to the molecular structure of the toxin. SNP rates and nonsynonymous and synonymous SNPs are indicated as in Fig. 2.**
deletion and truncated forms (9). Yet, our previous study demonstrated significant genetic diversity in tcdA and tcdB, consistent with the results of toxinotyping based on the restriction fragment length polymorphism (RFLP)-PCR method (19, 28). However, the simple classification of RFLP-PCR failed to explain more details about the epidemiological relevance and phylogenetic relationships between strains with different types of toxin genes, while the recently increased ST37 (RT017 [A15B13]) contribution (48%) for CDI cases and its epidemic potentiality in Asia (11, 16) make it urgent to build a more finely described pattern to reveal the intrinsic epidemic characteristics. In the present study, we introduce a novel approach that is more discerning and accurate; for example, all of the ST37 isolates from different regions in our study contained the same sequences of tcdA and tcdB, consistent with its low genetic diversity and the possible international spread. As an old, virulent A−B+ strain, 8864 (B16) has a variant 5 end similar to that of the hypervirulent A−B+ strain 8864 with exceptional pathogenicity (30, 31). And BG5, including the tcdB gene from ST37, has a high similarity with BG6 in the glucosyltransferase and autoprotease domains. These sequence similarities with the same confirmed highly pathogenic type indicate that BG6 may be the common ancestor of BG4 and BG5 and might be associated with the virulence characteristics of RT027 and ST37 strains, although other factors must also be involved (32, 33). For example, CD196 was not of clinical significance when it was isolated in 1985, despite having the same tcdA and tcdB sequences as R20291. Clades 2 to 6 are regarded as an “old” cluster, with clade 1 representing a relatively new cluster. So, we hypothesize that A−B+ strains evolved from an ancestral A−B+ strain after deletion of the 3′ end of the tcdA gene, consistent with that of RT017 (ST37) strains shown by whole-genome analysis to occupy a distinct lineage (24).

All of the ST37 isolates from different regions in our study contained the same sequences of tcdA and tcdB, consistent with its low genetic diversity and the possible international spread. As an old, virulent A−B+ strain (30, 31), 8864 (B16) has a variant 5′ end similar to those of common A−B+ strains (B13 to B15) and a variant 3′ end similar to that of the hypervirulent A−B+ RT027 strain (B12). This unusual arrangement may reflect a connection between RT027 and A−B+ strains and highlights the potential epidemic threat of an A−B+ strain in China. Another possibility is that the increased pathogenicity of toxin B enhances the virulence and spread of A−B+ strains. Some of the ST37 strains derive from the 1980s. The fact that their toxin genes have not changed in 30 years is quite remarkable and may be due to the inherent success of this strain.

This is the first systematic analysis of tcdA and tcdB in C. difficile.

FIG 3 Phylogenetic analyses of tcdA and tcdB. The Clostridium sordellii cytotoxin gene (GenBank accession number X82638) was used to root the phylogenetic trees of tcdA (left) and tcdB (right). Clades are represented by different colors, and A−B+ clades are marked with asterisks. Clade 1, as a new lineage, is distinguished from the older cluster, comprising clade 2 to clade 6, which is marked by a gray square.
from different regions of China, and it provides us a complete map of these two major toxin genes. This accurate map provides us not only the possible typing sites for nucleotide detection, but also the potential connection of prevalent strains among the world, such as the ST37 threat in China. Together with other genetic typing tools, such as MLST, the result could provide more accurate toxinic orientation, relationship maps, and interpretive classifications, which will facilitate epidemic risk evaluation beyond different period, population, and geographical spans to specified strains and improve CDI prevention and control.

ACKNOWLEDGMENTS

This work was supported by the National Natural Science Foundation of China (grant no. 81101218, 81201322, and 81301402). We thank Yuanyuan Zhang, Haiyin Wang, and Fenguian Li from the National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, for the technical support, and we thank Haihu Huang from Huashan Hospital, Shanghai, China, for generously providing samples.

We declare that we have no conflicts of interest.

REFERENCES


