Fourteen-genome comparison identifies DNA markers for severe-disease-associated strains of Clostridium difficile

Vincenzo Forgetta  
McGill University

Matthew T. Oughton  
Sir Mortimer B. Davis Jewish General Hospital

Pascale Marquis  
McGill University

Ivan Brukner  
Sir Mortimer B. Davis Jewish General Hospital

Ruth Blanchette  
Sir Mortimer B. Davis Jewish General Hospital

See next page for additional authors

Follow this and additional works at: http://digitalcommons.wustl.edu/open_access_pubs

Recommended Citation

Forgetta, Vincenzo; Oughton, Matthew T.; Marquis, Pascale; Brukner, Ivan; Blanchette, Ruth; Hauh, Kevin; Magrini, Vince; Mardis, Elaine R.; Gerding, Dale N.; Loo, Vivian G.; Miller, Mark A.; Mulvey, Michael R.; Rupnik, Maja; Dascal, Andre; and Dewar, Ken, "Fourteen-genome comparison identifies DNA markers for severe-disease-associated strains of Clostridium difficile." Journal of Clinical Microbiology.49,6. 2230-2238. (2011).  
http://digitalcommons.wustl.edu/open_access_pubs/2370

This Open Access Publication is brought to you for free and open access by Digital Commons@Becker. It has been accepted for inclusion in Open Access Publications by an authorized administrator of Digital Commons@Becker. For more information, please contact engeszer@wustl.edu.
Fourteen-Genome Comparison Identifies DNA Markers for Severe-Disease-Associated Strains of Clostridium difficile

Vincenzo Forgetta, Matthew T. Oughton, Pascale Marquis, Ivan Brukner, Ruth Blanchette, Kevin Haub, Vince Magrini, Elaine R. Mardis, Dale N. Gerding, Vivian G. Loo, Mark A. Miller, Michael R. Mulvey, Maja Rupnik, Andre Dascal and Ken Dewar

Published Ahead of Print 20 April 2011.

Updated information and services can be found at: http://jcm.asm.org/content/49/6/2230

These include:

REFERENCES
This article cites 61 articles, 28 of which can be accessed free at: http://jcm.asm.org/content/49/6/2230#ref-list-1

CONTENT ALERTS
Receive: RSS Feeds, eTOCs, free email alerts (when new articles cite this article), more»

Information about commercial reprint orders: http://journals.asm.org/site/misc/reprints.xhtml
To subscribe to another ASM Journal go to: http://journals.asm.org/site/subscriptions/
Fourteen-Genome Comparison Identifies DNA Markers for Severe-Disease-Associated Strains of Clostridium difficile *†

Vincenzo Forletta,1 Matthew T. Oughton,2 Pascale Marquis,3 Ivan Brukner,2 Kevin Haub,4 Vince Magrini,4 Elaine R. Mardis,4 Dale N. Gerding,5 Vivian G. Loo,6 Mark A. Miller,2 Michael R. Mulvey,7 Maja Rupnik,8 Andre Dascal,2 and Ken Dewar,1,3*

McGill University, Department of Human Genetics, Montreal, Canada; Sir Mortimer B. Davis Jewish General Hospital, Montreal, Canada; McGill University and Génome Québec Innovation Center, Montreal, Canada; The Genome Center at Washington University School of Medicine, St. Louis, Missouri; Infectious Disease Section and Research Service, Department of Medicine, Hines Veterans Affairs Hospital, and Loyola University Chicago Stritch School of Medicine, Hines, Illinois; McGill University Health Centre, Montreal, Canada; National Microbiology Laboratory, Winnipeg, Manitoba, Canada; and Institute of Public Health Maribor and University of Maribor, Faculty of Medicine, Maribor, Slovenia.

Published ahead of print on 20 April 2011.

Clostridium difficile is a common cause of infectious diarrhea in hospitalized patients. A severe and increased incidence of C. difficile infection (CDI) is associated predominantly with the NAP1 strain; however, the existence of other severe-disease-associated (SDA) strains and the extensive genetic diversity across C. difficile complicate reliable detection and diagnosis. Comparative genome analysis of 14 sequenced genomes, including one of a subset of NAP1 isolates, allowed the assessment of genetic diversity within and between strain types to identify DNA markers that are associated with severe disease. Comparative genome analysis of 14 isolates, including five publicly available strains, revealed that C. difficile has a core genome of 3.4 Mb, comprising ~3,000 genes. Analysis of the core genome identified candidate DNA markers that were subsequently evaluated using a multistrain panel of 177 isolates, representing more than 50 pulsosvars and 8 toxinotypes. A subset of 117 isolates from the panel had associated patient data that allowed assessment of an association between the DNA markers and severe CDI. We identified 20 candidate DNA markers for species-wide detection and 10,683 single nucleotide polymorphisms (SNPs) associated with the predominant SDA strain (NAP1). A species-wide detection candidate marker, the sspA gene, was found to be the same across 177 sequenced isolates and lacked significant similarity to those of other species. Candidate SNPs in genes CD1269 and CD1265 were found to associate more closely with disease severity than currently used diagnostic markers, as they were also present in the toxin A-negative and B-positive (A-B+) strain types. The genetic markers identified illustrate the potential of comparative genomics for the discovery of diagnostic DNA-based targets that are species specific or associated with multiple SDA strains.

* Corresponding author. Mailing address: McGill University, 740 Avenue Dr. Penfield, Room 7214, Montreal, Quebec, Canada H3A 1A4. Phone: (514) 398-3311, ext. 00089. Fax: (514) 398-1268. E-mail: ken.dewar@mcgill.ca.
† Published ahead of print on 20 April 2011.
‡ The authors have paid a fee to allow immediate free access to this article.

American pulse field type 1 (NAP1), ribotype 027, toxinotype III, or restriction-endonuclease type BI (36), which we refer to here as NAP1. Outbreaks have also been associated with other SDA strains, such as the NAP7/toxinotype V/ribotype 078 (NAP7) strain found in cases of human and animal disease (17, 38) and multiple toxin A-negative B-positive (A-B+) pulsotypes and ribotypes responsible for CDI outbreaks in Ireland, the United Kingdom, the United States, and Canada (1, 58).

To date, the monitoring of C. difficile infection in a hospital setting has been a reactive process in which patients are tested after symptoms emerge, usually by employing methods that are time-consuming and expensive and/or lack sufficient sensitivity (11, 26). Improved tests are sought in which all at-risk patients can be tested in a rapid, reliable, and cost-effective manner (45). To address this need, DNA-based diagnostic tests have been previously developed (2, 55, 56, 63), but they rely primarily on targets from previously known genomic regions, such as the genes tcdC (55, 63) and gyrA and gyrB (56). The available tests include numerous commercially available assays (Xpect toxin A/B test, from Remel, Inc., Lenexa, KS; BD GeneOhm Cd assay, from BD Diagnostics, San Diego, CA; ProGastro Cd assay, from Prodesse Inc., Waukesha, WI; Cepheid Xpert
C. difficile, from Cepheid, Sunnyvale, CA; and illumogene C. difficile, from Meridian Bioscience, Inc.).

With the advent of massively parallel DNA sequencing (MPS) technology (35), bacterial whole-genome sequencing has become rapid and affordable, and there are now sequences from well over 5,000 completed or in-progress microbial genomes available in public databases such as NCBI Entrez. MPS-based genome sequencing has been used in studies of C. difficile and other human pathogens, including a comparative genome analysis of 25 isolates that was performed to provide insight into the molecular evolution of C. difficile (20, 53) and a study of genome comparisons of 3 isolates that was used to identify potential virulence mechanisms in the NAP1 strain (59). These studies have confirmed the mobile nature of the C. difficile genome (54) and that genetic diversity among strains is high. In several studies (20, 22, 53), it has been shown that the NAP7-NAP8 strain type is highly divergent from other strain types and that the C. difficile core genome may be composed of only ~1,000 genes (20, 22, 53, 58). To date, MPS and comparative genome analyses of C. difficile have not been applied to the search for additional DNA-based diagnostic targets, as has been done for other human pathogens (6, 14, 15, 28).

In this report, we describe the comparative analysis of the genomes of 14 isolates of C. difficile. The genomes of 9 isolates, including 6 eastern Canadian and 3 reference isolates, were sequenced as part of this study, and the genome sequences from 5 additional publicly available isolates were also used in the analyses. Our main objective was to identify DNA targets for use in potentially addressing two major clinical issues: (i) determining whether a given patient is infected with C. difficile and, if so, (ii) determining whether the patient is infected with a strain associated with severe disease. We thus identified DNA-based diagnostic sequences that could be used to detect any isolate of C. difficile, as well as targets able to discriminate between SDA and non-SDA strain types. We also used comparative genome analysis to study the genetic diversity between strain types, estimate the size of the C. difficile core genome, and begin to investigate the existence of additional loci responsible for virulence. Candidate targets identified in the 14-genome analysis were reconfirmed with a larger panel of 177 isolates. Clinical records available for 117 of these isolates were cross-referenced by the use of target alleles to ascertain their association with disease severity.

**RESULTS**

**Genome sequencing and analysis.** We performed whole-genome sequencing and assembly for nine isolates of C. difficile (Table 1). Six isolates were of the predominant SDA strain type (NAP1), and collection dates ranged from 1984 to 2007. The international CIP 107932 reference isolate (isolated in 1984) and the BI-1 isolate (isolated in Minnesota in 1988 [47]) represented NAP1 isolates predating the CDI epidemic of 2003 to 2007. The other four NAP1 isolates (QCD-32g58, QCD-66c26, QCD-37x79, and QCD-97b34) were collected during the CDI epidemic from three locations across Canada. In addition to the six NAP1 isolates, we sequenced and assembled QCD-23m63, an SDA NAP7/toxinotype V/ribotype 078 (NAP7) isolate collected in 2007. The two non-NAP1 isolates sequenced were international reference strain VPI10463 (ATCC 43255) and a NAP2 isolate (QCD-63j42), collected in 1980 and 2005, respectively.

We also used the publicly available genome assemblies of another five isolates, including two additional NAP1 isolates (R20291 and CD196 [59]), NAP7 and NAP8 isolates (Human Microbiome Project [HMP]), and strain 630 (54) (Table 1). Genome assembly details and the NCBI and GenBank accession numbers for all 14 isolates are given in Table 2.

The nine genomes sequenced in this study ranged in size from 3.94 Mb (QCD-23m63) to 4.44 Mb (QCD-63j42) (Table 2), which corresponds to the range of genome sizes (3.90 to 4.29 Mb) observed in other C. difficile genome sequencing projects (Table 2). We generated 9 draft genome assemblies in which the contig numbers in the assemblies ranged from 16
However, the largest assembled contigs were 350 kb for all isolates, and the contig N80 calculation indicated that 80% of the assembled genomes was found in contigs of 59 kb (BI-1) to contigs of 300 kb (QCD-32g58). Given that C. difficile has a typical bacterial gene density (80 to 85%) and typical bacterial gene lengths (500 to 2,000 nucleotides [nt]), our genome assemblies provided contigs that could support synteny analyses combining

<table>
<thead>
<tr>
<th>Table 1. Characteristics of C. difficile isolates used in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Isolate</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>QCD-66c26</td>
</tr>
<tr>
<td>QCD-32g58</td>
</tr>
<tr>
<td>BI-1</td>
</tr>
<tr>
<td>CIP 107932</td>
</tr>
<tr>
<td>QCD-37x79</td>
</tr>
<tr>
<td>QCD-97b34</td>
</tr>
<tr>
<td>QCD-63q42</td>
</tr>
<tr>
<td>VPI 10463</td>
</tr>
</tbody>
</table>

Publicly available isolates

<table>
<thead>
<tr>
<th><strong>Isolate</strong></th>
<th><strong>Yr</strong></th>
<th><strong>PFGE type</strong></th>
<th><strong>Location</strong></th>
<th><strong>Source</strong></th>
<th><strong>Additional characteristic(s)</strong></th>
<th><strong>Genome size (Mb)</strong></th>
<th><strong>Contig no.</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>CD196</td>
<td>1985</td>
<td>NAP1</td>
<td>Paris, France</td>
<td>Nonepidemic strain</td>
<td></td>
<td>4.11</td>
<td>1</td>
</tr>
<tr>
<td>R20291</td>
<td>2006</td>
<td>NAP1</td>
<td>Stoke Mandeville, England</td>
<td>Outbreak associated</td>
<td></td>
<td>4.19</td>
<td>1</td>
</tr>
<tr>
<td>Strain 630</td>
<td>1980</td>
<td>NAP1</td>
<td>Zurich, Switzerland</td>
<td>CDI and PMC</td>
<td></td>
<td>4.29</td>
<td>2</td>
</tr>
<tr>
<td>NAP07</td>
<td>2008</td>
<td>NAP7</td>
<td>Unknown</td>
<td>Human feces</td>
<td></td>
<td>3.90</td>
<td>33</td>
</tr>
<tr>
<td>NAP08</td>
<td>2008</td>
<td>NAP8</td>
<td>Unknown</td>
<td>Human feces</td>
<td></td>
<td>4.68</td>
<td>24</td>
</tr>
</tbody>
</table>

(QCD-32g58) to 66 (BI-1). However, the largest assembled contigs were >350 kb for all isolates, and the contig N80 calculation indicated that 80% of the assembled genomes was found in contigs of >59 kb (BI-1) to contigs of >300 kb (QCD-32g58). Given that C. difficile has a typical bacterial gene density (80 to 85%) and typical bacterial gene lengths (500 to 2,000 nucleotides [nt]), our genome assemblies provided contigs that could support synteny analyses combining

<table>
<thead>
<tr>
<th>Table 2. Characteristics of C. difficile genome assemblies used in this study</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Isolate</strong></td>
</tr>
<tr>
<td>----------------</td>
</tr>
<tr>
<td>QCD-66c26</td>
</tr>
<tr>
<td>QCD-32g58</td>
</tr>
<tr>
<td>BI-1</td>
</tr>
<tr>
<td>CIP 1079324</td>
</tr>
<tr>
<td>QCD-37x79</td>
</tr>
<tr>
<td>QCD-97b34</td>
</tr>
<tr>
<td>QCD-63q42</td>
</tr>
<tr>
<td>VPI 10463</td>
</tr>
<tr>
<td>QCD-23m63</td>
</tr>
</tbody>
</table>

Publicly available isolates

<table>
<thead>
<tr>
<th><strong>Isolate</strong></th>
<th><strong>Technology</strong></th>
<th><strong>Genome assembly status</strong></th>
<th><strong>Genome size (nt)</strong></th>
<th><strong>No. of contigs</strong></th>
<th><strong>Largest contig (kb)</strong></th>
<th><strong>Contig N80 (kb)</strong></th>
<th><strong>NCBI or GenBank accession no.</strong></th>
<th><strong>Reference</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>CD196</td>
<td>GS-FLX</td>
<td>Complete</td>
<td>4,110,554</td>
<td>1</td>
<td>4,110.5</td>
<td>4,110.5</td>
<td>NZ_AAML000000000</td>
<td>This study</td>
</tr>
<tr>
<td>R20291</td>
<td>GS-20</td>
<td>Complete</td>
<td>4,191,339</td>
<td>1</td>
<td>4,191.3</td>
<td>4,191.3</td>
<td>NZ_ABHE000000000</td>
<td>This study</td>
</tr>
<tr>
<td>Strain 630</td>
<td>GS-FLX</td>
<td>Complete</td>
<td>4,290,252</td>
<td>1</td>
<td>4,290.2</td>
<td>4,290.2</td>
<td>NC_009089</td>
<td>Sebaihia et al. (54)</td>
</tr>
<tr>
<td>NAP07</td>
<td>GS-FLX</td>
<td>Draft</td>
<td>3,862,058</td>
<td>100</td>
<td>269.2</td>
<td>35.7</td>
<td>NZ_ADVM000000000</td>
<td>Human Microbiome Project</td>
</tr>
<tr>
<td>NAP08</td>
<td>GS-FLX</td>
<td>Draft</td>
<td>4,022,033</td>
<td>111</td>
<td>169.7</td>
<td>32.3</td>
<td>NZ_ADNX000000000</td>
<td>Human Microbiome Project</td>
</tr>
</tbody>
</table>

a nt, number of nucleotides.

b http://www.hmpdacc.org/.
sequence similarities in the context of information on gene order and orientation. Draft assemblies from the Human Microbiome Project also provided contigs of lengths (contig N80/H11022 30 kb) useful for gene annotation and the derivation of local synteny relationships.

We used whole-genome alignments to determine that a core of 3.4 Mb of orthologous genomic sequence was present in all of the 14 genomes. Within the core genome, the NAP7-NAP8 group of isolates tended to be the most divergent of all the groups, with an average percent identity of 97% compared to the NAP1 group, although this included regions of lower percent identity embedded within the syntenic regions (Fig. 1).

Polymorphism discovery. Within the 3.4-Mb core genome, we identified 127,442 single nucleotide polymorphisms (SNPs). Although other types of genomic variation, including insertions and deletions, were present, they were not analyzed in this study, as they accounted for less than 3% (3,211) of all instances of nucleotide-level variations. A phylogenetic tree constructed using the SNPs clustered the isolates into three distinct groups: the eight NAP1 isolates, the three NAP7-NAP8 isolates, and the three remaining isolates (ATCC 43255, 630, and QCD-63q42), which we refer here to as the R group (Fig. 2a). Levels of genetic variations among these three groups are greater than the level of variation within any one group (Fig. 2b). The NAP1 isolates differ from the NAP7-NAP8 group at 104,853 SNP positions and differ from the three remaining isolates (R group) at 17,076 SNP positions. The NAP7-NAP8 isolates differ from the R group of isolates at 96,302 SNP positions.

The members of the NAP1 group consisting of 8 isolates are highly identical within the core genome, with identities for any two members ranging from 9 to 62 polymorphic SNP positions (Fig. 2b). We identified 11 nonsynonymous nucleotide substitutions that distinguished members of a subset of Canadian NAP1 isolates that consists of QCD-32g58 (isolated in Quebec in 2004), QCD-66c26 (Quebec, 2007), and QCD-37x79 (Ontario, 2005) as well as the United Kingdom outbreak strain R20291 (Stoke Mandeville) from the others, namely, QCD-97b34 (Newfoundland, Canada, 2004), BI-1 (Minnesota, 1988), CIP107932 (France, 1984), and CD196 (France, 1985). This set of 11 SNPs includes the previously described mutation responsible for resistance to fluoroquinolones (10).

The members of the group of three NAP7-NAP8 isolates are also highly similar to each other, with at most 1,851 SNPs separating QCD-23m63 (NAP 7) from the Human Micro-

FIG. 1. Percent identity plot (top) and dot plot (bottom) depicting the whole-genome pairwise alignments of a NAP1 isolate (QCD-66c26) and a NAP7 isolate (QCD-23m63). The dot plot in the bottom panel depicts the colinearity of the two genomes, and the percent identity plot in the upper panel depicts the levels of nucleotide-level similarity of the two genomes (the red line indicates average percent identity).

FIG. 2. (a) Phylogenetic tree of 14 C. difficile genomes constructed using SNP data. Branches corresponding to partitions reproduced in less than 50% of replicate bootstrap determinations are collapsed. The percentages of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches. Genomes cluster into 3 distinct groups: NAP1 isolates (red), NAP7-NAP8 isolates (blue), and the 3 remaining isolates (black [R group]). (b) Numbers of polymorphic SNPs observed for the three groups (large boxes), as well as variations observed between isolates within each group (remaining values). SNPs (text at center bottom) that uniquely identify the members of each group are also indicated.
biome Project (HMP) NAP07 isolate. However, the two HMP isolates are very similar, with only 297 SNPs. The number of variations distinguishing the three isolates in the R group is greater than the number of variations within the other two groups, with each of the three isolates in the group having over 11,000 SNPs.

Genetic variations consistently present in members of a single group are candidate DNA markers of potential diagnostic value. We found 10,683 SNPs that distinguish the NAP1 group from the other two groups (Fig. 2b). There were also 89,954 SNPs that separated the NAP7-NAP8 set from all others and 6,342 SNPs that separated the remainder (R group) from the NAP1 and NAP7-NAP8 groups.

The 10,683 SNPs that are specific to the NAP1 group have a genome-wide distribution (Fig. 3a) and include a prominent cluster overlapping the tcdB cytotoxic gene (Fig. 3a). Among the other prominent clusters, 545 SNPs were observed in a 5.8-kb region of 6 genes (CD1265 to CD1270), annotated as a two-component system (TCS) adjacent to an ABC transporter (TCS-ABC) (Fig. 3b). The predicted protein sequences across this TCS-ABC locus showed no premature stop codons, providing preliminary evidence that these remain functional genes despite the increased level of genetic variation. Other prominent clusters include an ABC transporter (CD0336 to CD0337), and a hypothetical protein adjacent to a transcriptional regulator (CD3143 to CD3144).

Identification and validation of targets for SDA strains. To investigate the potential of NAP1 SNPs in the TCS-ABC locus as diagnostic targets associated with disease severity, we selected 2 candidate genes, CD1265 and CD1269, for comparison to 9 putative or known virulence genes (tcdA, tcdB, tcdC, tcdE, codY, fliC, groEL, gyrB, and mviN). These genes were PCR amplified and resequenced using a panel of 177 isolates of C. difficile. Isolates were selected to assess genetic variation for each candidate locus within the NAP1 strain and between strains of various PFGE types. The 177 isolates were collected predominantly from Canadian hospital, provincial, and national reference collections and chosen to sample variations in PFGE types. The isolates were also selected to encompass different collection sites and years of isolation. A second set of international strains facilitated another survey of C. difficile diversity based on toxinotyping (52). A total of 20 contributing sites were represented, including 17 Canadian, 2 American, and one European institution. There were 163 isolates collected in or after 2002, with 22 isolated prior to 2002. Sixty-seven isolates were classified as NAP1, 12 as NAP2, and 109 into more than 50 other pulsovars. Ten isolates were from the reference panel of toxinotyping strains, and 50 were from 9 hospitals in the province of Quebec, Canada. A total of 117 isolates were obtained from The Canadian Nosocomial Infection Surveillance Program (CNISP) and had available patient data. Seventy of the 117 clinical records reported severe CDI, which was defined as death associated with CDI or intensive care units or colectomy due to CDI (all measured at least 30 days after diagnosis).

Our set included 10 A-B+ isolates, all of which were of
PFGE type 00065 (NAP9; personal communication) and correlated with toxinotype VIII (22). In the patient data available for 100 of the isolates that were resequenced, SNP genotypes in two genes (CD1269 and CD1265) were seen to be associated with 70% (44/63) of severe-disease cases (comprising the three PFGE types NAP1, NAP7, and NAP9), whereas analysis by existing typing methods, such as PFGE NAP1 classification or determination of deletions in \( \text{tcdC} \), accounted for 52% (33/63) or 62% (39/63), respectively, and was limited to two PFGE types (NAP1 and NAP7). For example, the G allele for the SNP at genomic location 1,387,666 in CD1269 (Fig. 4) was associated with more severe-disease cases than the single base deletion at position 117 (\( \text{H}9004 \)\( 117 \)) in \( \text{tcdC} \), as revealed by the observation that the A-B+ strains have the G allele but lack the \( \text{tcdC} \) deletion. In contrast, analysis focusing on the \( \Delta 117 \) deletion in \( \text{tcdC} \) detected 33 of 63 severe-disease cases, and the larger deletions of at least 18 bp in the 3' end of \( \text{tcdC} \) accounted for 39 of 63 of severe-disease cases. Deletions in \( \text{tcdC} \) were diagnostic markers for two of three strain types previously associated with severe disease (NAP1 and NAP7), whereas analysis of the genotypes for each of 18 SNPs in \( \text{CD1269} \) (1/18) and \( \text{CD1265} \) (17/18) (the TCS-ABC locus) detected 44 of 63 severe-disease cases and specific alleles were associated with three SDA strain types (NAP1, NAP7, and A-B+ [NAP9/toxinotype VIII]) (Fig. 4). In addition, 5 of the SNPs in CD1265 are triallelic and partition the SDA strains into two groups (NAP1 versus NAP7 and A-B+ [NAP9/toxinotype VIII]) (Fig. 4). For example, the SNP at genomic location 1,383,226 has three alleles, C, T, and A. The A allele is associated with the NAP1 strain, the T allele with the NAP7 and A-B+ strains, and the C allele with all the others.

**Identification and validation of targets for species-level detection.** We selected a set of 12 genes (\( \text{spoIIAG} \), CD0596, CD0117, CD3014, CD0279, CD1795, CD2251, CD0017, \( \text{cdd1} \), \( \text{srlE} \), \( \text{prdB} \), and \( \text{sspA} \)) that displayed a high level of nucleotide identity across all 14 sequenced genomes as species-wide detection candidates (Table 3). After the panel of 177 isolates was resequenced, \( \text{sspA} \) was found to be 100% conserved at the nucleotide level for all isolates analyzed (Table 3). The remaining 11 candidates contained a few polymorphisms and did not deviate substantially from what was originally observed in the 14-genome analysis (Table 3). To further investigate whether these gene sequences could be used as DNA-based markers for specific detection of \( \text{C. difficile} \) (regardless of strain type) and yet remain specific to \( \text{C. difficile} \) and to no other species, we aligned the genomic sequence from each candidate gene re-
region to sequences in the NCBI nonredundant DNA database. 

Database hits for all 12 gene regions were well below 90% identity and/or 90% query coverage (Table 3). Furthermore, experiments performed using C. difficile PCR primers for all 12 genes did not lead to amplification products from Clostridium spiroforme and Mycobacterium intracellulare (data not shown).

### DISCUSSION

The primary objectives of this study was to identify DNA markers for C. difficile that could be used to test stool samples of patients and to determine (i) whether they are infected with C. difficile and, if they are infected, (ii) whether the particular strain is associated with severe disease. Our strain selection, which was influenced by the availability of isolates in well-characterized hospital collections, such as the Canadian Nosocomial Infection Surveillance Program, included an emphasis on isolates of the predominant epidemic strain (NAP1) as well as an isolate of another SDA strain (NAP7) previously reported in the literature (17, 38). Our strain choices comprised the majority of SDA strain types (21, 32, 46, 61) as well as two widely used reference research isolates (CIP107932 and VPI10463). Our testing of an extended panel of isolates, from an even wider diversity of strain types than those described here, further demonstrated that we have analyzed C. difficile representing a wide spectrum of naturally occurring genetic diversity.

In agreement with previous studies, we observed variations in genome size that are largely attributable to mobile genetic element activity (20, 53, 54, 59). Mobile genetic elements in C. difficile have been found to carry virulence factors (59); therefore, a more detailed comparative genome analysis of the NAP1 isolates from this study and others might provide further insight into the differential virulence characteristics observed within this strain. We observed 3.4 Mb of conserved sequence present in all 14 genomes, comprising 3,063 genes, a determination that differs substantially from observations in other studies, which estimated a smaller core genome of less than 1,000 genes (20, 22, 53, 58). We do not believe that this is a reflection of strain choices, as our most divergent pairs (NAP1 versus NAP7) were also recognized in other studies as being the most highly divergent (20). Rather, we believe the differences in interpretations of core genome size are due to differences in methodologies and analysis techniques. Given the high quality of the 3 completed genomes and 11 draft genomes used in our analyses, we were able to combine sequence comparisons and syntenic relationships to determine gene orthologs and observed that, compared to NAP1 references, the members of the NAP7-NAP8 group displayed a higher level of genetic diversity, with an average of 97% identity. We believe that in comparative genome hybridization (CGH) studies, for example, where probe mismatches can affect hybridization (41, 49) and where NAP1 or strain 630 genomes have been used as the reference, even a very low number of mismatches between the probe and target DNA can increase the false-negative rate (34, 58). This leads us to suggest that the C. difficile CGH arrays, which were designed using either strain 630 (R group) or NAP1 isolate QCD-32g58, may produce numerous false-negative hybridizations when tested using more a more distant related strain (e.g., NAP7) and have resulted in the calculation of a smaller core genome size. Other whole-genome sequence-based studies have used differing analytical techniques to determine the core genome size. For example, one study determined the core genome to consist of 622 genes, but only after the researchers stringently considered the non-recombining genes in the genome (20). Another study estimated the core genome to be comprised of 947 genes (53); however, as it was a gene-centered analysis based on sequence identity without synteny evaluation, that study may have classified genes of lower-than-average sequence identity, such as those present in the NAP7-NAP8 group, as nonorthologous. Our whole-genome multiple-alignment-based approach allowed the identification of orthologous regions of lower sequence identity which might not have been detectable using CGH arrays or might have been classified as below similarity thresholds by the use of a gene-centered approach.

At this time, our catalogue of genetic variation does not include insertions or deletions. Future work may reveal insertions or deletions in SDA strains in addition to those previously identified in tcdC (5, 32, 33, 36). Indeed, whereas deletions in tcdC have previously been hypothesized to lead to increased toxin production (33, 57, 62), findings from a recent study (39) indicated that deletions in tcdC do not predict the biological activity of the PaLoc toxin genes, providing further
motivation to identify all sources of genetic variation within the C. difficile genomes that may correlate with disease severity.

There are 64 SNPs that discriminate isolates within the NAP1 strain type, including 14 SNPs that separate eastern Canadian isolates from the others, and these demonstrate the potential of massively parallel sequencing for identification of SNPs suitable for subsequent intraspecies typing and tracking. The SNPs that distinguish the NAP1 strain from all other strains show an uneven genome-wide distribution and cluster in known pathogenicity genes as well as in genes with currently unrecognized roles in CDI. The other genes with clusters of NAP1 SNPs include a general stress gene (CD2599), a carbon starvation gene (CD2600), and an ABC transporter. The most prominent cluster of SNPs that discriminate the epidemic strain was observed in a genetic locus consisting of a two-component system (CD1269 and CD1270) adjacent to an ABC transporter (CD1265 to CD1268). The adjacency of these two loci has been shown to be functionally relevant in Bacillus and is present in other low-GC-content Gram-positive bacteria, including C. difficile (25). The function of the TCS-ABC system has been investigated previously in Bacillus subtilis and includes detoxification of antimicrobial compounds (42). The function of any particular TCS-ABC system is primarily dictated by the protein domains present in the histidine kinase gene (25). Sequence analysis of the histidine kinase gene (CD1270) in this TCS-ABC system suggests that it plays a role in detoxification (data not shown). In silico analysis indicates that this locus is not comprised of pseudogenes and thus may have a functional role. Future experiments are needed to confirm the expression of genes in the TCS-ABC system and investigate their roles in CDI.

Enzyme-based strategies for the detection of C. difficile have limited sensitivity and specificity (11). To address this, alternative gene targets, such as the tpi housekeeping gene, have been investigated (8). While this assay is specific to C. difficile, it requires the additional procedure of restriction fragment length polymorphism (RFLP) to achieve its specificity, making it less suitable in a clinical setting. Our genome analysis has accounted for additional severe-disease cases. We thank Manon Lorange for the isolation of C. difficile OGD-63q42. We also thank Marcel Behr for providing C. spiroforme and Mycobacterium intracellulare control DNAs. We thank the Human Microbiome Project (HMP) and the HMP DACC for prepublication data release of the C. difficile NAP07 (ADVM0000000000.1) and NAP08 (ADNX000000000.1) sequences.

This study was funded by Genome Canada and Genome Quebec (K.D.) and the NHGRI (E.R.M.). V.F. was the recipient of a Canadian Institutes of Health Research Doctoral Research Award. M.T.O. was the recipient of an AMMI Canada/CHIR/CFID/Bayer Healthcare research fellowship.

ACKNOWLEDGMENTS

We thank the DNA sequencing teams at the Genome Center at Washington University School of Medicine and the McGill University and Genome Quebec Innovation Centre for the sequencing of the C. difficile isolates. We thank Manon Lorange for the isolation of C. difficile OGD-63q42. We also thank Marcel Behr for providing C. spiroforme and Mycobacterium intracellulare control DNAs. We thank the Human Microbiome Project (HMP) and the HMP DACC for prepublication data release of the C. difficile NAP07 (ADVM0000000000.1) and NAP08 (ADNX000000000.1) sequences.

This study was funded by Genome Canada and Genome Quebec (K.D.) and the NHGRI (E.R.M.). V.F. was the recipient of a Canadian Institutes of Health Research Doctoral Research Award. M.T.O. was the recipient of an AMMI Canada/CHIR/CFID/Bayer Healthcare research fellowship.

REFERENCES

automated sequencer traces using phred. 1 Accuracy assessment. Genome Res. 8:175–185.


44. Forrester et al. J. CLIN. MICROBIOL. 2238