Comparison of whole-genome sequence-based methods and PCR ribotyping for subtyping of Clostridioides difficile

Baktash, A.; Corver, J.; Harmanus, C.; Smits, W.K.; Fawley, W.; Wilcox, M.H.; ... ; Kuijper, E.J.

AMERICAN SOCIETY FOR MICROBIOLOGY | Journal of Clinical Microbiology®

BACTERIOLOGY

Comparison of Whole-Genome Sequence-Based Methods and PCR Ribotyping for Subtyping of *Clostridioides difficile*

A. Baktash,[a] J. Corver,[a] C. Harmanus,[a,b] W. K. Smits,[a] W. Fawley,[c] M. H. Wilcox,[d] N. Kumar,[e] D. W. Eyre,[f] A. Indra,[g] A. Mellmann,[h] E. J. Kuijper[a,b]

[a]Department of Medical Microbiology, Section Experimental Bacteriology, Leiden University Medical Center, Leiden, The Netherlands
[b]National Reference Laboratory for *Clostridioides difficile*, National Institute of Public Health and the Environment, Leiden University Medical Center, Leiden, The Netherlands
[c]National Infection Service, Public Health England, and University of Leeds, Leeds, United Kingdom
[d]Department of Microbiology, Leeds Teaching Hospitals and University of Leeds, Leeds, United Kingdom
[e]Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, United Kingdom
[f]Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, United Kingdom
[g]Paracelsus Medical University of Salzburg, Salzburg, Austria
[h]Institute of Hygiene, University Hospital Münster, and National Reference Center for *C. difficile*, Münster Branch, Münster, Germany

ABSTRACT *Clostridioides difficile* is the most common cause of antibiotic-associated gastrointestinal infections. Capillary electrophoresis (CE)-PCR ribotyping is currently the gold standard for *C. difficile* typing but lacks the discriminatory power to study transmission and outbreaks in detail. New molecular methods have the capacity to differentiate better and provide standardized and interlaboratory exchangeable data. Using a well-characterized collection of diverse strains (N = 630; 100 unique ribotypes [RTs]), we compared the discriminatory power of core genome multilocus sequence typing (cgMLST) (SeqSphere and EnteroBase), whole-genome MLST (wgMLST) (EnteroBase), and single-nucleotide polymorphism (SNP) analysis. A unique cgMLST profile (more than six allele differences) was observed in 82 of 100 RTs, indicating that cgMLST could distinguish most, but not all, RTs. Application of cgMLST in two outbreak settings with RT078 and RT181 (known to have low intra-RT allele differences) showed no distinction between outbreak and nonoutbreak strains in contrast to wgMLST and SNP analysis. We conclude that cgMLST has the potential to be an alternative to CE-PCR ribotyping. The method is reproducible, easy to standardize, and offers higher discrimination. However, adjusted cutoff thresholds and epidemiological data are necessary to recognize outbreaks of some specific RTs. We propose to use an allelic threshold of three alleles to identify outbreaks.

KEYWORDS *Clostridioides difficile*, whole-genome sequencing, typing methods, core-genome MLST, whole-genome MLST

Clostridioides difficile is a Gram-positive anaerobic bacterium that is associated with nosocomial gastrointestinal infection (1, 2). It is estimated that there were almost 500,000 patients with *C. difficile* infection (CDI) and around 29,000 deaths in the United States in 2011 (2). Individuals with CDI are an important source of *C. difficile* transmission in health care settings (2). Typing of *C. difficile* is necessary for infection control, epidemiology, and evaluation of treatment. Several methods are used for typing *C. difficile*, including capillary electrophoresis (CE)-PCR ribotyping (3, 4) and multilocus sequence typing (MLST) (5). CE-PCR ribotyping is currently the gold standard. However, it does not provide sufficient discriminatory power to distinguish related strains (6). Furthermore, for CE-PCR ribotyping, standardization, and interlaboratory comparisons are difficult to establish (7). In contrast, this is relatively simple for sequence-based methods like MLST, in which sequence types (ST) are assigned based

Editor Daniel J. Diekema, University of Iowa College of Medicine

Address correspondence to E. J. Kuijper, e.j.kuijper@lumc.nl.

The authors declare a conflict of interest. D.W.E.: lecture fees from Gilead, outside the submitted work; E.J.K.: unrestricted research grant from Vedanta Bioscience, Boston.

Received 10 August 2021
Returned for modification 11 September 2021
Accepted 22 November 2021
Accepted manuscript posted online 15 December 2021
Published 16 February 2022

on the allele combinations of a limited number of housekeeping genes (5). Previously, it has been shown that the *C. difficile* phylogenetic tree, based on MLST, consists of five major clades. The majority of STs cluster in MLST clade 1. Hypervirulent ribotype (RT) 027 (ST1) belongs to clade 2, whereas hypervirulent RT078 (ST11) belongs to clade 5, which is divergent from the other clades (5, 8).

In the case of a suspected outbreak, CE-PCR ribotyping can be used in combination with multilocus variable-number tandem repeat (VNTR) analysis (MLVA) for subtyping of strains belonging to one PCR RT (9). This combination of methods is usually sufficient to type strains and understand transmission events. However, these methods do not provide sufficient information about strain characteristics (e.g., possession of virulence and resistance genes) and possible treatment failures (relapse versus reinfection). The techniques are also less suitable to study transmission over longer time scales, as tandem repeats are unstable and can come and go. This also limits the use of MLVA to determine the role of symptomatic and asymptomatic patients in hospital-acquired CDI (10). Therefore, typing methods with more discriminatory power and preferably based on better standardized whole-genome sequencing (WGS) are urgently needed.

There are two commonly applied methods to identify genomic variations using WGS. Single-nucleotide polymorphism (SNP) analysis usually uses a reference genome and detects SNPs between the reference genome and the studied genome (11). SNP analysis provides the highest resolution, but it is relatively slow, requires extensive bioinformatic tools, and is difficult to standardize, and typing nomenclature is missing (10, 12, 13). The second approach is based on gene-by-gene allelic profiling of the core genome (a set of genes that are conserved across most, if not all genomes studied) (cgMLST) or whole-genome (wgMLST) (14). cgMLST provides high discriminatory power, is more rapid than SNP analysis, offers reasonably accurate reproducibility, is reference independent (12), and could be used as a typing method since the cgMLST scheme is maintained by a centralized database (15).

Currently there are several cg/wgMLST schemes available for *C. difficile*, both commercially and publicly. The first commercial platform is SeqSphere$^+$ software (Ridom GmbH, Germany) comprising a scheme (the cgMLST.org nomenclature server) using up to 2,147 core genes and 1,357 accessory genes out of 3,756 genes present in strain 630 (15). The second is BioNumerics (bioMérieux, France) with the cgMLST/wgMLST scheme developed by Applied-Maths, comprising 1,999 core genes and 6,713 accessory genes and several other genes associated with virulence, antimicrobial resistance and others from different *C. difficile* strains (16). In addition to these two commercial platforms, there is a publicly available cg/wgMLST scheme from EnteroBase (University of Warwick, UK) consisting of 2,556 genes for the cgMLST scheme and up to 13,763 genes for the wgMLST scheme (17). The cgMLST scheme of EnteroBase (EB cgMLST) is also available through the Center for Genomic Epidemiology (cgMLSTFinder 1.1; https://cge.cbs.dtu.dk/services/cgMLSTFinder/).

Several studies have been published on the application of cgMLST (12, 15–17). Most studies show that cgMLST is concordant with CE-PCR ribotyping, but only a restricted number of RTs were analyzed, and outbreaks were not included. Recently, Seth-Smith et al. (18) showed that cgMLST predicted 36 RTs using nearly 300 well-characterized clinical strains from Switzerland. However, some RT complexes (RT078/126) had a low number of genetic differences, whereas other RTs (e.g., RT023) were very diverse (18). An in-depth analysis of backward compatibility between sequence-based methods and PCR ribotyping has been provided in a recent study (19).

The aim of this study is to assess the concordance between cg/wgMLST and CE-PCR ribotyping using a collection of 630 *C. difficile* strains belonging to 100 unique RTs. We analyzed the performance of CE-PCR ribotyping, cgMLST, wgMLST, and SNP analysis by using multiple software programs (SeqSphere$^+$ and EnteroBase). A second aim was to determine the optimal threshold to distinguish clonal strains from nonclonal in an outbreak setting. Importantly, our study shows that a threshold of up to three targets/alleles is needed for *C. difficile* isolates that are highly likely to belong to the same clone.

MATERIALS AND METHODS

Sequence data. The NCBI sequence read archive (SRA) was searched at the start of this study for *C. difficile* sequencing runs, and this resulted in 4,845 sequencing runs. Only sequence data generated on Illumina sequencing platforms and with available RT metadata were selected. A random selection of overrepresented strains (e.g., RT027 and RT078) was included. This approach resulted in 609 sequence runs that were analyzed. In addition to downloaded strains from the NCBI database, we also included 21 strains that were recently sequenced at the Leiden University Medical Center (LUMC). This included 15 Greek RT181 outbreak strains that were already sequenced for a previous study (PRJEB36956) (Table S1) (20) and 6 strains from an outbreak in The Netherlands due to RT078. For sequencing of strains, total DNA was isolated from cultured bacteria. A few colonies were emulsified in Tris/EDTA (TE) buffer and heated at 100°C for 10 min according to the protocol of Griffiths et al. (5). Chromosomal DNA was isolated using the QiaAmp blood and tissue kit (Qiagen) according to the instructions of the manufacturer. DNA was sequenced at Genome Scan B.V. (Leiden, The Netherlands) on an Illumina NovaSeq 6000 after preparation with the NebNext Ultra II DNA library prep kit for Illumina. This produced on average 3 million paired-end reads (read size 150 bp) per sample, with a minimum of 90% reads with a quality of 30 or more.

Ridom SeqSphere$^+$ cgMLST. Ridom SeqSphere$^+$ (version 6.0.2; Ridom GmbH, Münster, Germany) was run with default settings for quality trimming, *de novo* assembly, and allele calling on a Microsoft Windows operating system. Quality trimming occurred at both 5′ and 3′ ends until an average base quality of 30 was reached (length of 20 bases and a 120-fold coverage) (14, 15). *De novo* assembly was performed using the SKESA assembler version 2.3.0 (21) integrated in SeqSphere$^+$ (22) using default settings for SKESA. SeqSphere$^+$ scanned for the defined genes using BLAST (23) with the criteria described previously (14, 24). For further analysis, distance matrices, minimum spanning trees, and neighbor joining trees were constructed using the integrated features within SeqSphere$^+$ with the "pairwise ignoring missing values" option turned on.

EnteroBase cgMLST and wgMLST. cgMLST was performed using cgMLST Finder 1.1, available through the Center for Genomic Epidemiology (cgMLSTFinder 1.1; https://cge.cbs.dtu.dk/services/cgMLSTFinder/). Genomic data were processed using automated pipelines inside EnteroBase, as described in detail previously (25). In short, *de novo* assembly of Illumina sequence reads was performed using Spades version 3.10 (26). In order to pass quality control, assemblies were needed to comply with the criteria described previously (17). BLASTn and UBLASTP were used to align assemblies to alleles. The EnteroBase module MLSType was used to assess allele numbers and cluster types (25). cgMLST Finder 1.1 provides a distance matrix for analysis. Distance matrices were used to calculate the mean intra- and interallelic distance between different CE-PCR RTs. For wgMLST analysis, an *ad hoc* scheme was used based on the wgMLST scheme from EnteroBase (EB wgMLST) (17, 27). This *ad hoc* scheme was integrated in Ridom SeqSphere (15). *De novo* assembly, allele calling and further analysis were carried out as mentioned previously (under SeqSphere$^+$ cgMLST).

SNP analysis. SNPs were identified as previously described (28) using CSI Phylogeny 1.4 (http://cge.cbs.dtu.dk/services/CSIPhylogeny/). Default settings were used for the SNP analysis. *C. difficile* strain 630 (NC_009089) was used as the reference genome for the analysis of intra-RT SNP difference. The reference strain M120 (RT078) and a nonoutbreak related RT181 strain (obtained from a clinical isolate in 2019 in Romania) were used as a reference to analyze two CDI outbreaks with RT078 and RT181, respectively. In short, reads were mapped to the reference sequence using BWA (version 0.7.2) (29). Depth at each position was calculated using genomeCoverageBed, which is a component of BEDTools (version 2.16.2) (30). SNPs were called using mpileup, which is a component of SAMTools (version 0.1.18) (31). Mapping quality (minimum of 25 reads) and SNP quality (SNPs were filtered out if quality was below 30 or if they were called within the vicinity of 10 bp of another SNP) were calculated by BWA and SAMTools, respectively. CSIPhylogeny 1.4 provides a distance matrix for analysis. Distance matrices (based on pairwise comparison, missing data were excluded) were used to calculate the mean intra- and inter-RT SNP distance between different CE-PCR RTs.

Neighbor joining tree. A neighbor joining tree was constructed using integrated features with "pairwise ignoring missing values" option turned on in SeqSphere cgMLST. This analysis included 100 unique RTs. We selected strains from 75 different CE-PCR RTs of the standardized Leeds-Leiden collection (4). If not available, a random strain of a RT obtained from the sequencing runs from the SRA was used.

Mean intra-RT allele difference. Mean intra-RT allele difference and minimum and maximum range were determined for 19 RTs using distance matrices produced with cgMLST and wgMLST schemes and SNP analysis. These distance matrices showed the pairwise difference in alleles or in SNPs. From each RT, 3 to 13 strains were included. To prevent inclusion of related strains, e.g., from outbreak reports, we selected RTs with at least three random strains from different geographic locations and/or from different collection years.

Mean inter-RT allele difference. Mean inter-RT allele difference and minimum and maximum range were determined for 31 RTs using distance matrices produced with SeqSphere$^+$ cgMLST scheme. From each RT, one to three strains were included and compared with all strains (N = 630) included in this study, excluding the strains of the analyzed RT. To prevent inclusion of related strains, we selected random strains from different geographic locations and/or from different collection years.

Data availability. All genome sequence data generated as part of this study were submitted to the NCBI/ENA under study number PRJEB46469. The SRA accession numbers for the other analyzed genomes are provided in Table S1.

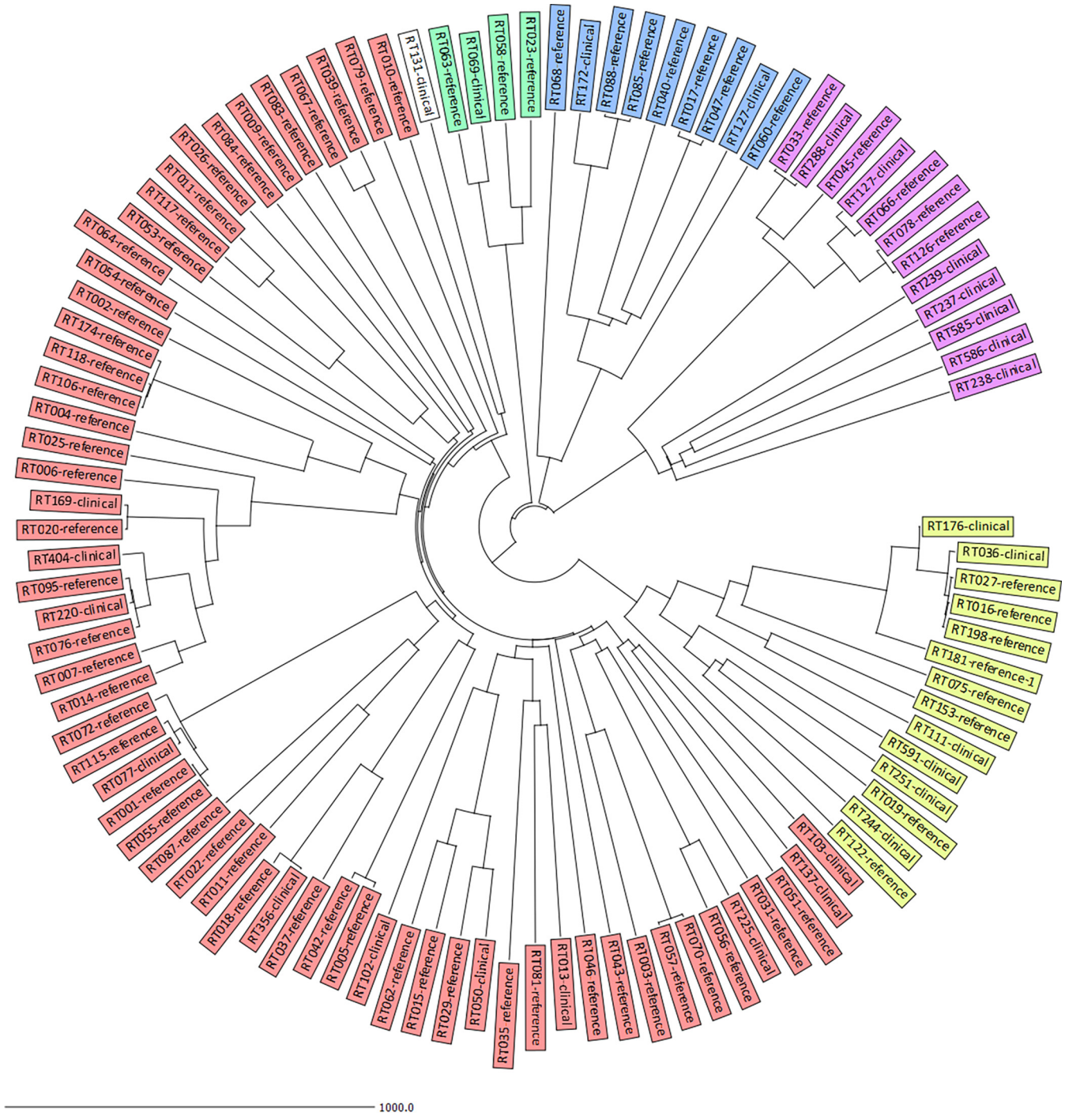

FIG 1 Neighbor joining tree from 100 unique ribotypes (RTs) based on SeqSphere cgMLST allele difference. Each RT is depicted with "RT*n*" followed by "reference" (belonging to the Leeds-Leiden collection) or "clinical" (non-Leeds-Leiden strain). RTs from multilocus sequence typing (MLST) clades 1, 2, 3, 4, and 5 are colored red, yellow, green, blue, and purple, respectively. RT131 has no designated MLST clade and is shown in white. The distance is given as the absolute allelic difference.

RESULTS

Ridom cgMLST can distinguish 82% of CE-PCR RTs. To test the concordance of cgMLST (SeqSphere$^+$) with CE-PCR ribotyping, we compared cgMLST and CE-PCR ribotyping using a selection of sequenced *C. difficile* strains with known RTs (10). Fig. 1 depicts a neighbor joining tree based on the Ridom SeqSphere$^+$ cgMLST scheme, including 100 different PCR RTs from all five MLST clades. Most RTs show a unique allelic profile in cgMLST. However, there are RTs within every MLST clade that show

low allele difference (six alleles or less) with other RTs, indicating that these RTs have likely recently evolved and cannot be distinguished with cgMLST using a simple allelic differences threshold.

When all included strains (n = 630 strains) from 100 unique RTs were analyzed with SeqSphere$^+$ cgMLST (Table 1), 82 RTs were distinguishable, i.e., the strains within these RTs differed by more than six alleles from strains within other RTs. Eighteen RTs (18%) from MLST clades 1, 2, and 5 clustered together with one to three other RTs from the same clade and had at most six allele differences. In Fig. 2, we show the RTs in each cluster and how these clusters vary at different thresholds (zero to six allele differences) to explore whether a lower threshold could distinguish different RTs. When the threshold was lowered from six to zero, the number of different RTs that clustered decreased from thirteen to two. Even at a threshold of zero allele differences, RT045 and RT127 remained clustered, demonstrating the limitation of short-read sequencing and cgMLST as only able to capture part of the genetic diversity present.

Intra-RT allele differences vary per RT and per MLST clade. We determined the mean number of allele differences between strains from the same RT to see whether it varies by RT and tested if intra-RT allele differences vary between MLST clades. We also compared the mean intra-RT allele or SNP differences with cgMLST, wgMLST, and SNP analysis. Mean intra-RT allele difference varied between RTs (Fig. 3A and Table 2). The method with the smallest scheme (SeqSphere$^+$ cgMLST) showed the lowest intra-RT allele difference average (mean range of 5 to 376 alleles), whereas SNP analysis showed the highest average (mean range of 67 to 2,563 SNPs). Fig. 3A also shows that the so-called hypervirulent RT027 (clade 2) had intra-RT allele differences of 8.4 (SeqSphere$^+$ cgMLST), 10.7 (EB cgMLST), 18.1 (EB wgMLST), and 100.7 (SNP). Another frequently found hypervirulent RT, RT078 (clade 5), showed allele differences of 13.2, 15.5, 29.3, and 139.4, respectively. The most frequently found RT in Europe, RT014 (clade 1), showed allele differences of 148.1, 173, 258.8, and 855.7, respectively. EB wgMLST and SNP analysis showed similar results as cgMLST but showed much higher average intra-RT allele and SNP differences. The RT with lowest intra-RT allele difference for clade 1 was RT002 (64 cgMLST alleles and 140 SNPs), and the highest was RT056 (650 alleles and 2,563 SNPs). The RT with the lowest intra-RT difference from clade 2 was RT181 (11 alleles and 67 SNPs), whereas the highest was RT036 (39 alleles and 120 SNPs). Lastly, RT126 from clade 5 showed the lowest difference (18 allele and 130 SNP differences) and RT127 the highest (379 allele and 592 SNP differences). SNP analysis showed the highest resolution and often more than 2 times more differences in comparison with wgMLST.

To determine the applicability and the extent of background diversity for outbreak studies of a given clade, including involving a novel RT, we analyzed the observed variety in allele and SNP difference for clades 1, 2, and 5. The average intraclade allele difference was calculated by combining the averages per RT within a clade (Fig. 3B). Clade 1 had the highest average allele differences for SeqSphere$^+$ cgMLST, EB cgMLST, EB wgMLST, and SNP analysis (114, 136, and 171 allele difference and 685 SNPs, respectively). Followed by clade 5 with 39, 49, and 66 allele differences and 177 SNPs, respectively. Clade 2 had the lowest average intra-RT allele difference (9, 12, and 18 allele differences and 100 SNPs, respectively).

Inter-RT allele differences vary by MLST clades. We determined the mean number of allele differences between strains from different RTs and analyzed with SeqSphere$^+$ cgMLST whether inter-RT allele difference varies by clade. Mean inter-RT allele difference varied between RTs (Fig. S1A to E). Comparing all MLST clade sequences to RT014 (in clade 1) showed a mean inter-RT allele difference of 1,787.9 (range, 0 to 2,213 alleles). In clade 2, the mean inter-RT allele difference comparing to RT027 was 1,890.2 (range, 3 to 2,214). Compared to all MLST clades, RT078 in clade 5 showed a mean inter-RT of 1,781.2 (range, 3 to 2,214 alleles). In contrast, for RTs 023 (clade 3) and 017 (clade 4), there were higher mean inter-RT allele differences 2,088.3 (range, 450 to 2,193) and 2,124.3 (range, 14 to 2,207 alleles), respectively.

Clade 5 had the lowest mean number of inter-RT allele differences (Fig. S1F), 1,627.2 alleles (range, 0 to 2,014). The first quartile (25% of the data were below this

TABLE 1 Clustering between PCR RTs[a]

Threshold (in alleles)	Studied PCR RT		Matching other PCR RT strain		
	RT	No. of strains[b]	RT	No. of strains[b]	Clade
6	020	1/20	076	1/2	1
	016	1/1	027	5/23	2
			036	1/4	
			176	4/16	
	027	3/23	036	2/4	2
		10/23	176	13/16	
		2/23	198	1/2	
	036	1/4	176	1/16	2
	033	2/46	288	2/2	5
	045	2/15	078	16/58	5
		2/15	126	7/29	
	066	1/2	078	3/58	5
		1/2	126	1/29	
	078	39/58	126	23/29	5
5	018	1/18	356	1/13	1
	016	1/1	027	2/23	2
			176	1/16	
			198	1/2	
	027	4/23	036	1/4	2
		10/23	176	6/16	
		2/23	198	1/2	
	036	1/4	176	2/16	2
	033	1/46	288	1/2	5
	045	2/15	078	7/58	5
		2/15	126	4/29	
		3/15	127	2/17	
	066	1/2	078	4/58	5
		1/2	126	2/29	
	078	31/58	126	21/29	5
4	016	1/1	027	1/23	
	027	6/23	036	1/4	2
		9/23	176	5/16	
	033	2/46	288	2/2	
	045	2/15	078	5/58	
		2/15	126	4/29	
		3/15	127	3/17	
	066	1/2	078	1/58	
	078	25/58	126	15/29	
3	018	1/18	356	3/13	1
	027	3/23	036	1/4	2
		6/23	176	3/16	2
	045	1/15	078	1/58	
		1/15	126	1/29	
		3/15	127	3/17	
	078	18/58	126	13/29	
2	001	1/14	055	1/1	1
	018	1/18	356	3/13	1
	016	1/1	027	1/23	
	027	3/23	176	2/16	
	045	2/15	126	2/29	
		1/15	127	2/17	
	078	8/58	126	4/29	
1	018	1/18	356	6/13	1
	045	1/15	127	1/17	
0	045	2/15	127	2/17	

[a]The clustering between PCR ribotypes (RTs) is shown only in one direction; e.g., the comparison between RT016 and RT027 at threshold 6 is shown only in the RT016 row and not again in the RT027 row.
[b]Number of strains that cluster with another PCR RT.

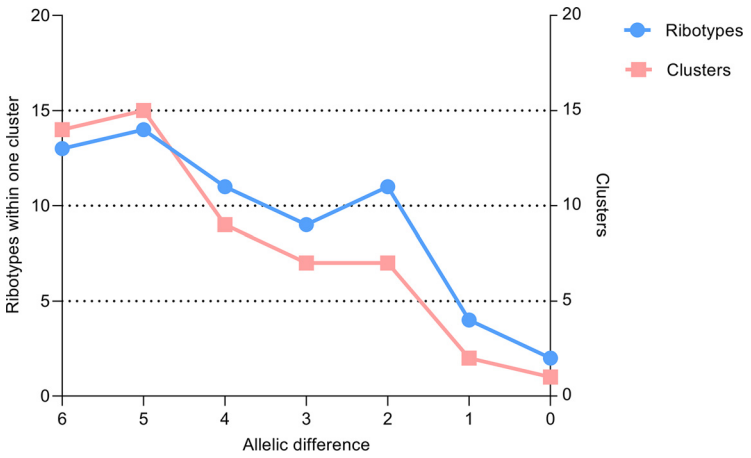

FIG 2 Clustering of different PCR RTs at different thresholds using SeqSphere$^+$ core genome multilocus sequence typing (cgMLST; zero to six allelic difference). The number of clustering RTs is shown in blue, and the number of clusters at every threshold is shown in pink; e.g., at three allele differences, nine different RTs belong to seven clusters.

point) of the samples showed 398 allele differences. This meant that the RT in clade 5 were more related to other RTs from clade 5. In contrast, clade 3 showed the highest inter-RT allele difference of 2,088.3 (range, 289 to 2,195) with a first quartile of 2,070 alleles. This indicates that RTs belonging to clade 3 are less related to each other and to RTs of other clades. When all clades are combined, the mean inter-RT allele difference was 1,742.1 (range, 0 to 2,215).

WGS-based typing methods may not distinguish outbreak strains from nonoutbreak strains in RTs with a low intra-RT allele differences. CE-PCR ribotyping has a low resolution in comparison with WGS-based typing for outbreak analysis. However, even with the increased resolution of WGS-based typing, it remains crucial to understand what defines an outbreak. The current proposed threshold for cgMLST for isolates that are likely to belong to the same clone is six alleles or fewer (15). To assess this interpretative threshold, we compared cgMLST, wgMLST, and SNP analysis in two outbreak settings. We selected outbreak strains from MLST clades 2 (RT181) and 5 (RT078) (Fig. 4A and B), since both clades have a low average allele differences between strains in contrast to strains from MLST clade 1. A non-outbreak-related RT181 strain and *C. difficile* strain M120 (RT078), respectively, were used as reference strains for the genetic analysis. Outbreak strains were defined as having a well-established epidemiological link (e.g., nursed in the same ward) combined with at most 6 allele differences with cgMLST. Control strains belonged to similar CE-PCR RTs as the outbreaks strains or to other CE-PCR RTs from the same clade. We analyzed the distance matrices of two clusters containing confirmed outbreaks and nonoutbreak strains with cgMLST, wgMLST, and SNP analysis.

The first CDI suspected outbreak we analyzed was due to RT078 (clade 5) in a Dutch general hospital, involving six patients in the gastroenterology ward between October-December 2018 (Fig. 4A). Three of these cases (outbreak cases in red circles) were epidemiologically linked by location and onset of disease. The other three cases with RT078 CDI (green circles) were admitted 1 month later. Twelve additional control samples from clade 5 were added to this collection. These included five Leeds-Leiden reference strains (RT033, RT045, RT066, RT078, and RT126) (4) and seven other strains (RT045, RT066, RT126, RT127, and RT078 [N = 3]). Fig. 4A depicts the minimum-spanning tree (based on SeqSphere$^+$ cgMLST) of the studied isolates. Three clusters (six alleles or less) could be recognized, each comprising epidemiologically related and unrelated strains. The three outbreak cases showed a clustering and had 0 allele differences.

A report on the second outbreak has been published recently (20) and therefore is not described again in detail. This outbreak occurred in a Greek 180-bed rehabilitation

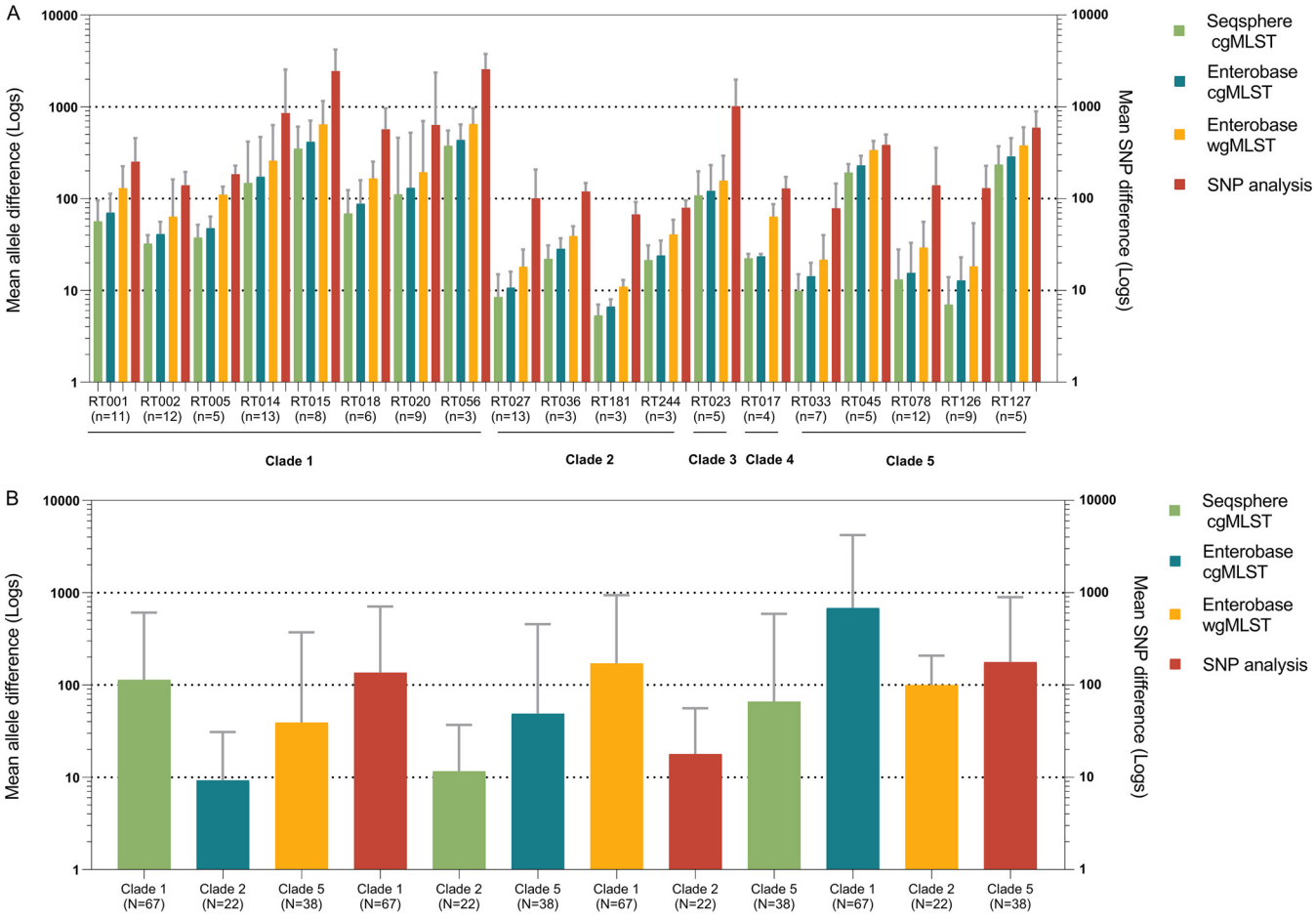

FIG 3 (A) Mean intra-RT allele and single-nucleotide polymorphism (SNP) difference shown for individual RTs from MLST clade 1 (RT001 to RT056), clade 2 (RT027 to RT244), clade 3 (RT023), clade 4 (RT017), and clade 5 (RT033 to RT127). Mean intra-RT allele difference per RT is shown in light green, turquois, and orange for SeqSphere$^+$ core genome multilocus sequence typing (cgMLST), EnteroBase (EB) cgMLST1, and EB whole-genome multilocus sequence typing (wgMLST), respectively. Mean intra-RT SNP difference per RT is shown in red. (B) Mean intra-RT allele and SNP difference shown for MLST clade 1, clade 2, and clade 5. Mean intra-RT allele difference per clade is shown in light green, turquois, and orange for SeqSphere$^+$ cgMLST, EB cgMLST, and EB wgMLST, respectively. Mean intra-RT SNP difference per clade is shown in red.

clinic involving 15 CDI patients infected with RT181 (clade 2) at the orthopedics and neurological wards between March and April 2019 (Fig. 4B). All 15 patient isolates showed a minimal number of allele differences (0 to 2 alleles) to the most closely related case, with a range across all cases of zero to eight. Seven control samples from clade 2 were added to this collection, including Leeds-Leiden reference strains of RT016, RT027, and RT198; one strain of RT036 and RT176; and two strains of RT181. Fig. 4B shows the minimum-spanning tree based on SeqSphere$^+$ cgMLST. Two clusters could be recognized, each composed of epidemiologically related and unrelated strains. Cluster 1 contained both confirmed outbreak strains (RT181, N = 15) and one control strain of RT181. Therefore, the current threshold of 6 alleles or less incompletely separated the outbreak of RT181 from the background diversity.

The strains within each cluster of RT078 or RT181 were either labeled as outbreak strain or control strain. The distance matrices of both clusters were visualized in graphs (Fig. 5A and B) with each data point representing a distance in alleles or SNPs between two strains. We calculated the range of allele or SNP difference of outbreak (O) strains (range O) and compared it with the range of allele or SNP difference of nonoutbreak (NO) strains (range NO). The NO range depicts allele or SNP difference between the control strain and the outbreak strains. The area between the upper limit of range O and the lower limit of range NO determines the area where adjustment of the threshold is possible, provided that outbreak strains and nonoutbreak

TABLE 2 Several ribotypes with given mean intra-ribotype difference and minimum and maximum range in alleles or single-nucleotide polymorphisms (SNPs) per whole-genome sequence (WGS) method

Ribotype	WGS method	Mean	Minimum range	Maximum range
001 (N = 11)	SeqSphere cgMLST	56.7	9	96
	EB cgMLST	70.5	11	113
	EB wgMLST	130.1	17	226
	SNP analysis	251.7	115	445
002 (N = 12)	SeqSphere cgMLST	32.3	6	40
	EB cgMLST	41.1	9	56
	EB wgMLST	63.7	13	162
	SNP analysis	139.5	41	196
005 (N = 5)	SeqSphere cgMLST	37.8	19	52
	EB cgMLST	47.5	23	64
	EB wgMLST	110.5	46	135
	SNP analysis	184	133	229
014 (N = 13)	SeqSphere cgMLST	148.1	4	419
	EB cgMLST	173	6	470
	EB wgMLST	258.8	13	635
	SNP analysis	855.7	68	2,556
015 (N = 8)	SeqSphere cgMLST	349.7	14	610
	EB cgMLST	416.9	15	708
	EB wgMLST	643	26	1,159
	SNP analysis	2,456.2	59	4,206
018 (N = 6)	SeqSphere cgMLST	69.1	4	124
	EB cgMLST	87.7	6	159
	EB wgMLST	165.5	62	254
	SNP analysis	567.9	146	965
020 (N = 9)	SeqSphere cgMLST	111.4	4	461
	EB cgMLST	131.2	5	524
	EB wgMLST	193.2	6	701
	SNP analysis	632.5	23	2,368
056 (N = 3)	SeqSphere cgMLST	376	22	554
	EB cgMLST	436.3	24	643
	EB wgMLST	650.3	43	963
	SNP analysis	2,562.7	212	3,780
027 (N = 13)	SeqSphere cgMLST	8.4	2	15
	EB cgMLST	10.7	3	16
	EB wgMLST	18.1	7	28
	SNP analysis	100.7	18	208
036 (N = 3)	SeqSphere cgMLST	22	10	31
	EB cgMLST	28.3	15	37
	EB wgMLST	39	20	50
	SNP analysis	120	95	148
181 (N = 3)	SeqSphere cgMLST	5.3	3	7
	EB cgMLST	6.7	4	8
	EB wgMLST	11	10	13
	SNP analysis	67.3	47	92
244 (N = 3)	SeqSphere cgMLST	21.3	3	31
	EB cgMLST	24	5	35
	EB wgMLST	40.7	8	59
	SNP analysis	79.3	56	96
023 (N = 5)	SeqSphere cgMLST	108.7	24	199
	EB cgMLST	121.3	24	233
	EB wgMLST	157.5	41	293
	SNP analysis	1,014.7	140	1,980
017 (N = 4)	SeqSphere cgMLST	22.3	20	25
	EB cgMLST	23.5	21	25
	EB wgMLST	63.7	40	87
	SNP analysis	129.3	76	172
033 (N = 7)	SeqSphere cgMLST	9.9	0	15
	EB cgMLST	14.2	2	20
	EB wgMLST	21.6	1	40
	SNP analysis	78.7	25	146

(Continued on next page)

TABLE 2 (Continued)

Ribotype	WGS method	Mean	Minimum range	Maximum range
045 (N = 5)	SeqSphere cgMLST	192.2	15	239
	EB cgMLST	230.2	20	293
	EB wgMLST	337.8	30	425
	SNP analysis	384.8	52	499
078 (N = 12)	SeqSphere cgMLST	13.2	3	28
	EB cgMLST	15.5	5	33
	EB wgMLST	29.3	7	56
	SNP analysis	139.4	27	358
126 (N = 9)	SeqSphere cgMLST	7	0	14
	EB cgMLST	12.8	6	23
	EB wgMLST	18.3	7	54
	SNP analysis	129.7	41	228
127 (N = 5)	SeqSphere cgMLST	235.4	17	372
	EB cgMLST	288	21	455
	EB wgMLST	379.1	30	599
	SNP analysis	592.2	143	893

strains do not overlap. The larger the area, the better the method can discriminate between outbreak and nonoutbreak strains. Fig. 5A shows that all WGS-based methods could distinguish between confirmed outbreak and nonoutbreak RT078 strains, since there is no overlap between range O and range NO. SNP analysis had the best discriminatory power, followed by EB wgMLST and cgMLST, which showed the lowest discriminatory power. Fig. 5B shows that wgMLST and SNP analysis could discriminate between outbreak and nonoutbreak RT181 strains, whereas cgMLST showed overlap in their ranges. Ranges O and NO are shown in Table S2 for both clusters and all applied typing methods. No overlap was seen between range O and range NO from cluster 1 from the RT078 CDI outbreak. For SeqSphere cgMLST and EB cgMLST, cluster 1 showed a difference of three alleles and two alleles between range O and range NO, respectively. Furthermore, the difference between range O and range NO was 6 alleles for wgMLST and 8 SNPs for SNP analysis, indicating that the threshold could be lowered. However, cluster 1 from the RT181 CDI outbreak showed overlap between range O and range NO in cgMLST but not in wgMLST and SNP analysis, suggesting that the threshold could only be adjusted in wgMLST and SNP analysis.

DISCUSSION

CE-PCR ribotyping is currently the gold standard for typing *C. difficile* strains. This method is an indirect way to compare genomes of *C. difficile* strains since it is based on the lengths and numbers of ribosomal interspace regions between 16S and 23S rRNA and not on the sequence of this intergenic space (3). Therefore, hypothetically the CE-PCR RT banding pattern of two genetically unrelated strains can have an identical appearance. Likewise, similarity of two RTs does not necessarily predict genetic relatedness between strains. Our aim was to compare the discriminatory power of cgMLST (SeqSphere$^+$ and EnteroBase) and wgMLST (EnteroBase) with SNP analysis for typing *of C. difficile*. We also touched upon the backward compatibility of WGS-based methods with CE-PCR ribotyping, but our goal was not to fully study backward compatibility.

We tested the concordance between SeqSphere$^+$ cgMLST and CE-PCR ribotyping and found that 82 of 100 different PCR RTs had a unique cgMLST profile using a cutoff of at most six alleles differences. Certain strains with distinct RTs were indistinguishable by SeqSphere$^+$ cgMLST, similar to data from Seth-Smith et al. (18), who found genomes of different RTs (RT078/RT126, RT106/RT500) clustering with maximum of 9 allelic difference. In agreement with these findings, we found that RTs from clades 1 and 5 had the lowest mean inter-RT allele difference and was directly followed by clade 2. Our results are also consistent with data from others (17, 18, 32). Finally, Frentrup et

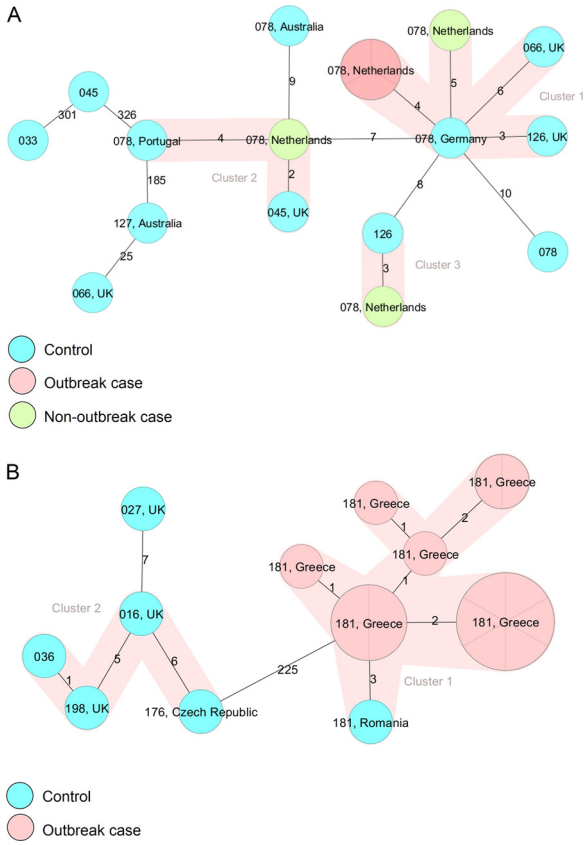

FIG 4 SeqSphere$^+$ cgMLST analysis with minimum-spanning trees of two suspected CDI outbreaks of RT078 and RT181. (A) Minimum-spanning tree of RT078 (clade 5) CDI suspected outbreak with six cases (RT078, shown in red and in green) of which three were confirmed (shown in largest septated red circle) and added control strains of RTs belonging to clade 5 (RT033, RT045, RT066, RT078, RT126, and RT127 shown in blue). (B) Minimum-spanning tree of RT181 (clade 2) CDI suspected outbreak with 15 suspected and 15 confirmed cases (RT181, shown in septated red circles) and added control strains of RTs belonging to clade 2 (RT016, RT027, RT036, RT176, RT181 and RT198 shown in blue). The size and septation of the circle in the minimum-spanning trees corresponds to the number of included strains. The numbers between each circle correspond to the number of different alleles between the strains. The colored shadowing of circles represents a cluster with at most six allele differences that are genetically related. One or more strains inside a circle means that these strains have zero allele differences.

al. (17), observed clustering of several RTs (e.g., RT001/RT241, RT106/RT500, and RT078/RT126) from MLST clades 1 and 5, also in agreement with our observations.

The genome clustering of various RTs was reduced, but not eliminated, by decreasing the threshold from 6 to 0 allele difference. The clustering between two strains of RT045 and two strains of RT127 at a threshold of 0 alleles in SeqSphere$^+$ cgMLST was verified with EB cgMLST and SNP analysis. With EB cgMLST, one clustering pair of RT045 and RT127 showed one allele difference, whereas the other remained at zero allele differences. Verification with SNP analysis showed two and seven SNP differences. This observation shows that cgMLST cannot predict all CE-PCR RTs and instead would require additional epidemiological information to analyze strains belonging to RT045 and RT127 together.

We showed that the mean inter-RT allele differences per clade were high using SeqSphere$^+$ cgMLST. This means that the genomes of most RTs differ (by more than 1,700 alleles) from other RTs. However, there are RTs that tend to cluster with related RTs (e.g., RT014, RT027, and RT078) and have lower inter-RT allele differences. Strains from clades 3 and 4 have higher mean inter-RT allele differences, indicating that the RTs within these clades are less related to each other. The inter-RT allele and SNP differences from EB cgMLST, wgMLST, and SNP analysis differed in scale but followed similar

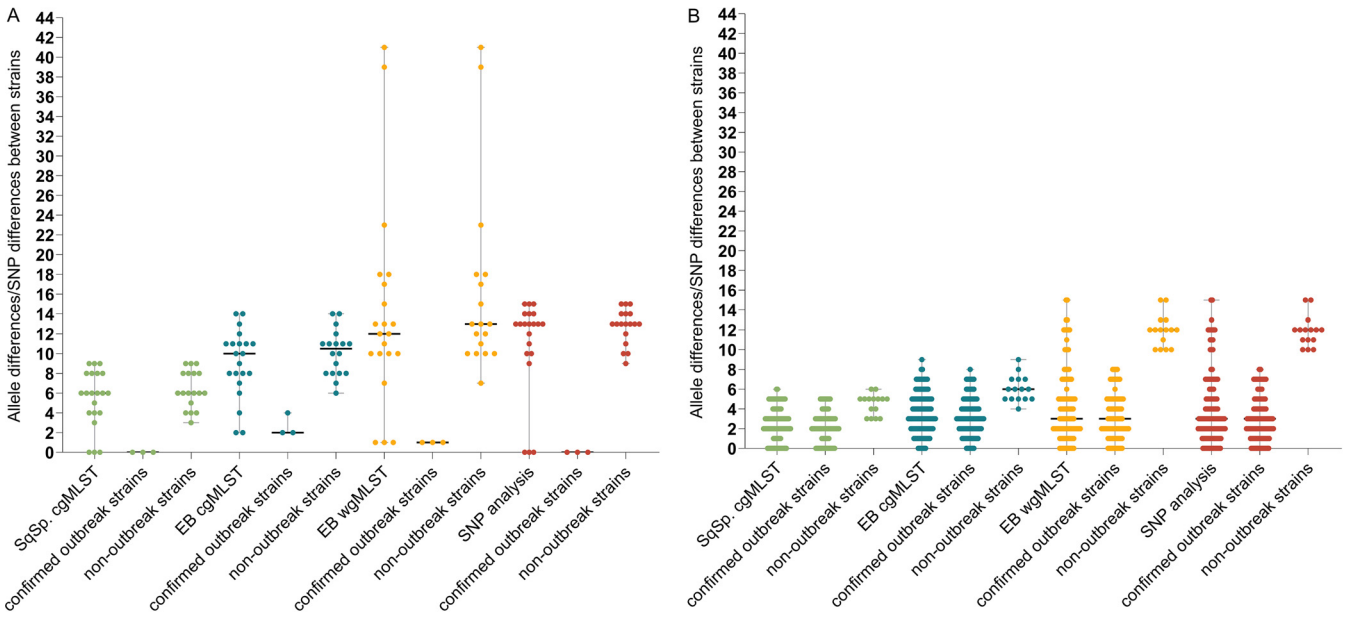

FIG 5 Visualized distance matrices of strain pairs based on cgMLST, wgMLST, and SNP analysis of isolates of cluster 1 as described in Fig. 4. (A) Visualized distance matrix of strain pairs belonging to cluster 1 of RT078. (B) Visualized distance matrix of strain pairs belonging to cluster 1 of RT181. Allele difference per pair of strains is shown in light green, turquois, and orange for SeqSphere$^+$ cgMLST, EB cgMLST, and wgMLST, respectively. SNP difference per pair of strains is shown in red.

patterns, reflecting the overall discriminatory power of each approach, i.e., the mean allele differences between strains from the same RTs with SeqSphere+ cgMLST and EB cgMLST are lower in comparison with EB wgMLST and SNP analysis.

Based on our observations in two CDI outbreaks, we conclude that cgMLST has less discriminatory power than wgMLST and SNP analysis in MLST clades with low intra-RT allele differences. Lower diversity in some strains may reflect recent emergence and dissemination and/or lower mutation rates, resulting in less diversity and therefore a lower intra-RT allele difference (33, 34). For outbreaks caused by RTs belonging to other clades than 2 and 5, the performance of cgMLST is comparable with SNP analysis, similar to what was found in other studies (15, 35). Based upon the Oxfordshire data set (33), Frentrup et al. had a similar conclusion regarding cgMLST and SNP analysis (17). They showed that *C. difficile* genomes that differ by zero, one, or two alleles generally also differ by at least two SNPs, using a logistic regression model, and concluded that cgMLST is equivalent to SNP analysis for identifying transmission chains between patients. Bletz et al. showed similar results between cgMLST and SNP analysis in detecting clusters when an outbreak due to ST1 was investigated (15). Therefore, we propose to apply a lower threshold of three alleles in comparison to the initially published threshold of six alleles (15) when using cgMLST in outbreak situations. In the study by Eyre et al., the evolutionary rate of *C. difficile* was estimated to be 0.74 SNVs (95% confidence interval, 0.22 to 1.40) per genome per year (36). They expected zero to two SNPs to occur when isolates are obtained less than 124 days apart and three SNPs when isolates were obtained 124 to 364 days apart. However, only vegetative *C. difficile* isolates obtained from patients were analyzed. According to Weller and Wu, sporulation reduces the evolutionary rate of *Firmicutes* (37). Therefore, we expect that the evolutionary rate of *C. difficile* is lower during CDI transmission than during CDI within a patient, since the spores need time to transmit to another patient and otherwise lie dormant in the surroundings in a health care facility or in the environment for a long period. Accordingly, we expect that outbreak strains will generally fall within zero to two alleles. Nevertheless, we recommend a threshold of three alleles to compensate for any assembly artifacts when less conservative pipelines are used and for outbreaks that last longer than 124 days (38). Even if we have optimized the threshold

based on our data presented here, by applying this threshold of three alleles, we still encountered difficulties with interpretation of the RT181 outbreak. Here, we observed a broad range (0 to 8 SNPs) within the outbreak strains. A possible explanation could be that there were multiple introductions of different RT181 strains to the hospital or that SNPs arose on several consecutive transmission events, which is not very unlikely since the outbreak had a prolonged course and was also affecting other hospitals. Since RT181 has very recently emerged, limited sequence data are available to assess the intra-RT allele difference more accurately. Also, this situation demonstrates that even with an optimized threshold, epidemiological information is always necessary and helpful to interpret ambiguous typing results.

The main strength of our study is that we compared the performance of several typing methods, in contrast to previous studies (12, 15–17). We also expanded the collection of *C. difficile* strains and tested more than 600 sequenced strains belonging to 100 unique RTs. Our study has also some limitations. The lack of sufficient available genome sequences from strains belonging to clades 3 and 4 limits the generalizability of our findings. Although the concordance of EB wgMLST with CE-PCR ribotyping was not tested, the results can be extrapolated from SeqSphere$^+$ cgMLST, EB cgMLST, and SNP analysis, since the discriminatory power of EB wgMLST lies between the latter two. We could not verify the correctness of the strain RTs, as we had only access to the information as deposited by the researchers. There are also a few RTs that have similar banding patterns and could be misidentified. The best example is the similarity of RT014 with RT020; they have an almost identical PCR banding pattern, but they differ substantially from each other by cgMLST. Though we only studied two outbreaks, we carefully selected the outbreaks by choosing RTs with low intra-RT allele variation. Finally, we have not tested long-read sequencing, from which *in silico* PCR ribotyping can theoretically also be obtained.

A concern with application of cgMLST is the availability of various cgMLST schemes and software programs. The centralized databases need resources to maintain their databases of sequentially numbered alleles. To tackle the problem of the need for a centralized database and to rapidly identify related genomes against a background of thousands of other identified genomes, hash-based cgMLST has been developed (12). It is based on cgMLST but converts alleles to a fixed unique hash or short string of letters. Whether traditional cgMLST or hash-based MLST is used, as Werner et al. proposed, it is favorable that a fixed cgMLST scheme is adopted to standardize comparisons (35). Furthermore, there are logistical and cost considerations for routine implementation of cgMLST. Reference laboratories are needed with a good infrastructure to sequence strains on a routine basis while keeping the costs in mind as well. Cost-wise, WGS-based methods are becoming competitive with the current gold-standard, CE-PCR ribotyping. In the near future, cgMLST could be used as the typing method, and when the discriminatory limit of cgMLST is reached (e.g., outbreak with strains with low intra-RT allele differences), wgMLST or SNP analysis could be applied.

In summary, cgMLST has the potential to replace CE-PCR ribotyping for *C. difficile*. The method provides similar differentiation of strains, is easy to standardize, is reproducible, and shows a high discriminatory power. Several cgMLST-based typing methods have emerged with all their specific advantages and disadvantages (12, 15, 17). For the time being, it remains unclear whether one method will get the preference over other methods or that every center will use its own method. A consensus group could be assembled to harmonize these efforts as has been done previously for CE-PCR ribotyping (4).

SUPPLEMENTAL MATERIAL

Supplemental material is available online only.

SUPPLEMENTAL FILE 1, PDF file, 2.8 MB.

ACKNOWLEDGMENTS

We thank B. V. H. Hornung for his assistance with data collection.

D.W.E. received lecture fees from Gilead, outside the submitted work. E.J.K. received an unrestricted research grant from Vedanta Bioscience (Boston, MA). The other authors declare no conflict of interest.

REFERENCES

1. Keller JJ, Kuijper EJ. 2015. Treatment of recurrent and severe *Clostridium difficile* infection. Annu Rev Med 66:373–386. https://doi.org/10.1146/annurev-med-070813-114317.
2. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. 2016. *Clostridium difficile* infection. Nat Rev Dis Primers 2:16020. https://doi.org/10.1038/nrdp.2016.20.
3. Indra A, Huhulescu S, Schneeweis M, Hasenberger P, Kernbichler S, Fiedler A, Wewalka G, Allerberger F, Kuijper EJ. 2008. Characterization of *Clostridium difficile* isolates using capillary gel electrophoresis-based PCR ribotyping. J Med Microbiol 57:1377–1382. https://doi.org/10.1099/jmm.0.47714-0.
4. Fawley WN, Knetsch CW, MacCannell DR, Harmanus C, Du T, Mulvey MR, Paulick A, Anderson L, Kuijper EJ, Wilcox MH. 2015. Development and validation of an internationally-standardized, high-resolution capillary gel-based electrophoresis PCR-ribotyping protocol for *Clostridium difficile*. PLoS One 10:e0118150. https://doi.org/10.1371/journal.pone.0118150.
5. Griffiths D, Fawley W, Kachrimanidou M, Bowden R, Crook DW, Fung R, Golubchik T, Harding RM, Jeffery KJ, Jolley KA, Kirton R, Peto TE, Rees G, Stoesser N, Vaughan A, Walker AS, Young BC, Wilcox M, Dingle KE. 2010. Multilocus sequence typing of *Clostridium difficile*. J Clin Microbiol 48:770–778. https://doi.org/10.1128/JCM.01796-09.
6. Knetsch CW, Connor TR, Mutreja A, van Dorp SM, Sanders IM, Browne HP, Harris D, Lipman L, Keessen EC, Corver J, Kuijper EJ, Lawley TD. 2014. Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011. Euro Surveill 19:20954.
7. Huber CA, Foster NF, Riley TV, Paterson DL. 2013. Challenges for standardization of *Clostridium difficile* typing methods. J Clin Microbiol 51:2810–2814. https://doi.org/10.1128/JCM.00143-13.
8. Dingle KE, Griffiths D, Didelot X, Evans J, Vaughan A, Kachrimanidou M, Stoesser N, Jolley KA, Golubchik T, Harding RM, Peto TE, Fawley W, Walker AS, Wilcox M, Crook DW. 2011. Clinical *Clostridium difficile*: clonality and pathogenicity locus diversity. PLoS One 6:e19993. https://doi.org/10.1371/journal.pone.0019993.
9. Knetsch CW, Lawley TD, Hensgens MP, Corver J, Wilcox MW, Kuijper EJ. 2013. Current application and future perspectives of molecular typing methods to study *Clostridium difficile* infections. Euro Surveill 18:20381. https://doi.org/10.2807/ese.18.04.20381-en.
10. Janezic S, Rupnik M. 2019. Development and implementation of whole genome sequencing-based typing schemes for *Clostridioides difficile*. Front Public Health 7:309. https://doi.org/10.3389/fpubh.2019.00309.
11. Eyre DW, Walker AS. 2013. *Clostridium difficile* surveillance: harnessing new technologies to control transmission. Expert Rev Anti Infect Ther 11:1193–1205. https://doi.org/10.1586/14787210.2013.845987.
12. Eyre DW, Peto TEA, Crook DW, Walker AS, Wilcox MH. 2019. Hash-based core genome multilocus sequence typing for *Clostridium difficile*. J Clin Microbiol 58:e01037-19. https://doi.org/10.1128/JCM.01037-19.
13. Bush SJ, Foster D, Eyre DW, Clark EL, De Maio N, Shaw LP, Stoesser N, Peto TEA, Crook DW, Walker AS. 2020. Genomic diversity affects the accuracy of bacterial single-nucleotide polymorphism-calling pipelines. Gigascience 9:giaa007. https://doi.org/10.1093/gigascience/giaa007.
14. Mellmann A, Andersen PS, Bletz S, Friedrich AW, Kohl TA, Lilje B, Niemann S, Prior K, Rossen JW, Harmsen D. 2017. High interlaboratory reproducibility and accuracy of next-generation-sequencing-based bacterial genotyping in a ring trial. J Clin Microbiol 55:908–913. https://doi.org/10.1128/JCM.02242-16.
15. Bletz S, Janezic S, Harmsen D, Rupnik M, Mellmann A. 2018. Defining and evaluating a core genome multilocus sequence typing scheme for genome-wide typing of *Clostridium difficile*. J Clin Microbiol 56:e01987-17. https://doi.org/10.1128/JCM.01987-17.
16. Gateau C, Deboscker S, Couturier J, Vogel T, Schmitt E, Muller J, Menard C, Turcan B, Zaidi RS, Youssouf A, Lavigne T, Barbut F. 2019. Local outbreak of *Clostridioides difficile* PCR-ribotype 018 investigated by multi locus variable number tandem repeat analysis, whole genome multi locus sequence typing and core genome single nucleotide polymorphism typing. Anaerobe 60:102087. https://doi.org/10.1016/j.anaerobe.2019.102087.
17. Frentrup M, Zhou Z, Steglich M, Meier-Kolthoff JP, Goker M, Riedel T, Bunk B, Sproer C, Overmann J, Blaschitz M, Indra A, von Muller L, Kohl TA, Niemann S, Seyboldt C, Klawonn F, Kumar N, Lawley TD, Garcia-Fernandez S, Canton R, Del Campo R, Zimmermann O, Gross U, Achtman M, Nubel U. 2020. A publicly accessible database for *Clostridioides difficile* genome sequences supports tracing of transmission chains and epidemics. Microb Genom 6:mgen000410.
18. Seth-Smith HMB, Biggel M, Roloff T, Hinic V, Bodmer T, Risch M, Casanova C, Widmer A, Sommerstein R, Marschall J, Tschudin-Sutter S, Egli A. 2021. Transition from PCR-ribotyping to whole genome sequencing based typing of *Clostridioides difficile*. Front Cell Infect Microbiol 11:681518. https://doi.org/10.3389/fcimb.2021.681518.
19. Moore MP, Wilcox MH, Walker AS, Eyre DW. 2021. K-mer based prediction of *Clostridioides difficile* relatedness and ribotypes. bioRxiv https://doi.org/10.1101/2021.05.17.444522:2021.05.17.444522.
20. Kachrimanidou M, Baktash A, Metallidis S, Tsachouridou O, Netsika F, Dimoglou D, Kassomenaki A, Mouza E, Haritonidou M, Kuijper E. 2020. An outbreak of *Clostridioides difficile* infections due to a 027-like PCR ribotype 181 in a rehabilitation centre: epidemiological and microbiological characteristics. Anaerobe 65:102252. https://doi.org/10.1016/j.anaerobe.2020.102252.
21. Souvorov A, Agarwala R, Lipman DJ. 2018. SKESA: strategic k-mer extension for scrupulous assemblies. Genome Biol 19:153. https://doi.org/10.1186/s13059-018-1540-z.
22. Junemann S, Sedlazeck FJ, Prior K, Albersmeier A, John U, Kalinowski J, Mellmann A, Goesmann A, von Haeseler A, Stoye J, Harmsen D. 2013. Updating benchtop sequencing performance comparison. Nat Biotechnol 31:294–296. https://doi.org/10.1038/nbt.2522.
23. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
24. Ruppitsch W, Pietzka A, Prior K, Bletz S, Fernandez HL, Allerberger F, Harmsen D, Mellmann A. 2015. Defining and evaluating a core genome multilocus sequence typing scheme for whole-genome sequence-based typing of *Listeria monocytogenes*. J Clin Microbiol 53:2869–2876. https://doi.org/10.1128/JCM.01193-15.
25. Zhou Z, Alikhan NF, Mohamed K, Fan Y, Agama Study G, Achtman M, Agama Study Group. 2020. The EnteroBase user's guide, with case studies on *Salmonella* transmissions, *Yersinia pestis* phylogeny, and *Escherichia* core genomic diversity. Genome Res 30:138–152. https://doi.org/10.1101/gr.251678.119.
26. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S, Prjibelski AD, Pyshkin AV, Sirotkin AV, Vyahhi N, Tesler G, Alekseyev MA, Pevzner PA. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol 19:455–477. https://doi.org/10.1089/cmb.2012.0021.
27. Jolley KA, Bray JE, Maiden MCJ. 2018. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res 3:124. https://doi.org/10.12688/wellcomeopenres.14826.1.
28. Kaas RS, Leekitcharoenphon P, Aarestrup FM, Lund O. 2014. Solving the problem of comparing whole bacterial genomes across different sequencing platforms. PLoS One 9:e104984. https://doi.org/10.1371/journal.pone.0104984.
29. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
30. Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842. https://doi.org/10.1093/bioinformatics/btq033.

31. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Genome Project Data Processing S. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352.
32. Miles-Jay A, Weissman SJ, Adler AL, Baseman JG, Zerr DM. 2021. Whole genome sequencing detects minimal clustering among *Escherichia coli* sequence type 131-H30 isolates collected from United States children's hospitals. J Pediatric Infect Dis Soc 10:183–187. https://doi.org/10.1093/jpids/piaa023.
33. Eyre DW, Fawley WN, Best EL, Griffiths D, Stoesser NE, Crook DW, Peto TE, Walker AS, Wilcox MH. 2013. Comparison of multilocus variable-number tandem-repeat analysis and whole-genome sequencing for investigation of *Clostridium difficile* transmission. J Clin Microbiol 51:4141–4149. https://doi.org/10.1128/JCM.01095-13.
34. Didelot X, Eyre DW, Cule M, Ip CL, Ansari MA, Griffiths D, Vaughan A, O'Connor L, Golubchik T, Batty EM, Piazza P, Wilson DJ, Bowden R, Donnelly PJ, Dingle KE, Wilcox M, Walker AS, Crook DW, Peto TE, Harding RM. 2012. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. Genome Biol 13:R118. https://doi.org/10.1186/gb-2012-13-12-r118.
35. Werner A, Molling P, Fagerstrom A, Dyrkell F, Arnellos D, Johansson K, Sundqvist M, Noren T. 2020. Whole genome sequencing of *Clostridioides difficile* PCR ribotype 046 suggests transmission between pigs and humans. PLoS One 15:e0244227. https://doi.org/10.1371/journal.pone.0244227.
36. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, O'Connor L, Ip CLC, Golubchik T, Batty EM, Finney JM, Wyllie DH, Didelot X, Piazza P, Bowden R, Dingle KE, Harding RM, Crook DW, Wilcox MH, Peto TEA, Walker AS. 2013. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. N Engl J Med 369:1195–1205. https://doi.org/10.1056/NEJMoa1216064.
37. Weller C, Wu M. 2015. A generation-time effect on the rate of molecular evolution in bacteria. Evolution 69:643–652. https://doi.org/10.1111/evo.12597.
38. Kuenzli AB, Burri S, Casanova C, Sommerstein R, Buetti N, Seth-Smith HMB, Bodmer T, Egli A, Marschall J. 2020. Successful management of a *Clostridioides difficile* ribotype 027 outbreak with a lean intervention bundle. J Hosp Infect 106:240–245. https://doi.org/10.1016/j.jhin.2020.07.034.