diagnostics

MDPI

Article

Retrospective Definition of *Clostridioides difficile* PCR Ribotypes on the Basis of Whole Genome Polymorphisms: A Proof of Principle Study

Manisha Goyal [1], Lysiane Hauben [2], Hannes Pouseele [3], Magali Jaillard [4], Katrien De Bruyne [2], Alex van Belkum [1,*] and Richard Goering [5]

[1] BioMérieux, Open Innovation and Partnerships, 3 Route du Port Michaud, 38390 La Balme Les Grottes, France; manisha.goyal@biomerieux.com

[2] BioMérieux, Applied Maths NV, 9830 Sint-Martens-Latem, Belgium; lysiane.hauben@biomerieux.com (L.H.); katrien.debruyne@biomerieux.com (K.D.B.)

[3] BioMérieux, Industry, 69290 Craponne, France; hannes.pouseele@biomerieux.com

[4] BioMérieux, 69280 Marcy l'Etoile, France; magali.dancette@biomerieux.com

[5] Department of Medical Microbiology and Immunology, Creighton University School of Medicine, 2500 California Plaza, Omaha, NE 68178, USA; richardgoering@creighton.edu

* Correspondence: alex.vanbelkum@biomerieux.com; Tel.: +33-609-487-905

Received: 17 November 2020; Accepted: 10 December 2020; Published: 12 December 2020

Abstract: *Clostridioides difficile* is a cause of health care-associated infections. The epidemiological study of *C. difficile* infection (CDI) traditionally involves PCR ribotyping. However, ribotyping will be increasingly replaced by whole genome sequencing (WGS). This implies that WGS types need correlation with classical ribotypes (RTs) in order to perform retrospective clinical studies. Here, we selected genomes of hyper-virulent *C. difficile* strains of RT001, RT017, RT027, RT078, and RT106 to try and identify new discriminatory markers using in silico ribotyping PCR and De Bruijn graph-based Genome Wide Association Studies (DBGWAS). First, in silico ribotyping PCR was performed using reference primer sequences and 30 *C. difficile* genomes of the five different RTs identified above. Second, discriminatory genomic markers were sought with DBGWAS using a set of 160 independent *C. difficile* genomes (14 ribotypes). RT-specific genetic polymorphisms were annotated and validated for their specificity and sensitivity against a larger dataset of 2425 *C. difficile* genomes covering 132 different RTs. In silico PCR ribotyping was unsuccessful due to non-specific or missing theoretical RT PCR fragments. More successfully, DBGWAS discovered a total of 47 new markers (13 in RT017, 12 in RT078, 9 in RT106, 7 in RT027, and 6 in RT001) with minimum q-values of 0 to 7.40 × 10^{-5}, indicating excellent marker selectivity. The specificity and sensitivity of individual markers ranged between 0.92 and 1.0 but increased to 1 by combining two markers, hence providing undisputed RT identification based on a single genome sequence. Markers were scattered throughout the *C. difficile* genome in intra- and intergenic regions. We propose here a set of new genomic polymorphisms that efficiently identify five hyper-virulent RTs utilizing WGS data only. Further studies need to show whether this initial proof-of-principle observation can be extended to all 600 existing RTs.

Keywords: ribotyping; *C. difficile*; unitigs; markers; WGS; GWAS; CDI; in silico PCR

1. Introduction

Clostridioides difficile (*C. difficile*), formerly known as *Clostridium difficile*, is an anaerobic, spore-forming Gram-positive bacterial species that can survive in harsh environments. It can withstand high temperatures, exposure to ultraviolet light, toxic chemicals, and exposure to antibiotics. Colonization by *C. difficile* is asymptomatic. The development of disease is mostly driven by host factors and disruption of the gut microbiome by frequent consumption of antibiotics [1–3]. Toxigenic strains of *C. difficile* can be a lethal cause of *C. difficile* infection (CDI), which is commonly associated with post antibiotic diarrhea [4,5]. *C. difficile* is present in the environment and can be transmitted to patients or healthcare workers through contact with contaminated surfaces. Inter-human spread mainly occurs through the fecal–oral route. *C. difficile* spores are intrinsically resistant to antibiotics and remain viable during antibiotic treatment. Clindamycin, cephalosporins, and fluoroquinolones are considered as major antibiotics associated with CDI [6]. Food or water contamination, gastric acid-suppression, and asymptomatic carriage in the community are the potential risk factors of community acquired CDI [7]. One-third of the total CDI burden occurring in the USA in 2011 was community-associated [8]. CDI caused half a million hospital-acquired infections and 29,000 deaths in 2012 in the United States [8] and approximately 40,000 cases of CDI in Europe [9]. Only a limited number of studies reported emerging CDI in Asia [10]. The increasing incidence of CDI and rapid evolution of antibiotic resistance in *C. difficile* has become a global threat to public health [11–13].

CDI diagnosis allows early pathogen isolation and treatment of infection, thereby reducing the potential of CDI transmission. Various diagnostic procedures for CDI are available, including toxigenic culture, cell cytotoxic neutralization assay, glutamate dehydrogenase assay, the detection of toxins by enzyme immunoassays, nucleic acid amplification-based molecular tests, etc. [14–19]. Still, epidemiological *C. difficile* strain typing is necessary to identify outbreaks within a hospital or the wider community and facilitates understanding of the dissemination of infections. Ribotyping is a classical technique for *C. difficile* typing initially based on hybridization patterns of conserved ribosomal RNA probe sequences [20,21]. Ribotype (RT) analysis has also been extremely important in the long-term surveillance of CDI [22]. While traditional typing methods such as restriction endonuclease analysis (REA) and pulsed-field gel electrophoresis (PFGE) were widely used in the past, PCR-based ribotyping is the current method of choice for *C. difficile* typing [23,24]. PCR ribotyping is dependent on the amplification of the intergenic spacer region (ISR) between 16S and 23S rRNA genes [25–28]. Since most bacterial species encode multiple ribosomal alleles in their genomes, multiple fragments of different lengths are amplified when different species but also different strains are considered [25,26]. There are still considerable constraints on PCR ribotyping including elevated costs, a higher probability of false-positive results, and a lack of interlaboratory portability [10,29,30].

Bacterial whole genome sequencing (WGS) has the potential to provide more detailed epidemiological information than classical PCR ribotyping [23,31]. To further explore a WGS-based approach to *C. difficile* typing, backward compatibility with PCR ribotyping is essential [32]. Previous studies reported the association of RT001, RT017, and RT027 with lethal CDI and considered those isolates as hyper-virulent [33,34]. A survey conducted in the North East of England concluded that RT001, RT027, and RT106 were among the most prevalent and dangerous clones [35]. In the United States and Europe, RT001, RT014, RT020, RT027, and RT078 have been identified as predominant [36,37]. RT017 is a globally emerging toxigenic RT and can be found on almost every continent [38,39]. Thus, here, we tested both in silico PCR ribotyping and the De Bruijn graph-based Genome Wide Association Study (DBGWAS) [40] for their capacity to perform retrospective PCR ribotyping for *C. difficile* RT001, RT017, RT027, RT078, and RT106. These strains were chosen as a test set representing global, long-term circulating and clinically relevant epidemic strains. The primary study goal was to develop a proof-of-principle procedure for sequence-based *C. difficile* strain typing with retrospective compatibility to established PCR RTs.

2. Materials and Methods

2.1. In Silico PCR-Based Ribotyping

We performed in silico PCR using canonical ribotyping PCR primers. Based on the reference sequences 16S-USA and 23S-USA (Table 1), in silico PCR was performed using the subsequence search tool in BioNumerics v7.6 software (Applied Maths NV, Sint Martens-Latem, Belgium). Besides 7 genomes obtained from Creighton University, 23 *C. difficile* genomes of five selected RTs were downloaded from NCBI to verify in silico amplification of the ISR region (Supplementary Table S1).

Table 1. Primer pair used for in silico PCR-based ribotyping of *Clostridioides difficile*.

Primer	Gene Target	GenBank Accession No.	Sequence (5′–3′)	Tm (°C)	Reference
16S-USA (Forward)	16S rRNA gene	FN545816	(12293)GTGCGGCTGGATCACCTCCT (12312)	71.0	Xiao et al., 2012 (46)
23S-USA (Reverse)	23S rRNA gene	FN545816	(12621)CCCTGCACCCTTAATAACTTGACC (12598)	67.1	

2.2. DBGWAS-Mediated Discovery New RT-Specific Markers

A total number of 160 *C. difficile* genome assemblies (training set) of 14 different RTs including hyper-virulent RT001, RT017, RT027, RT078, and RT106 were used for the discovery of unique RT genomic markers (Table 2). This small training set allowed for the development of discriminatory markers to characterize the five major RTs among the 14 different RTs. Genomes were collected from the National Center for Biotechnology and Information (NCBI; www.ncbi.nlm.nih.gov), Creighton University, and the Enterobase databases (https://enterobase.warwick.ac.uk). Metadata of these genomes are summarized in Supplementary Table S2.

Table 2. *C. difficile* ribotypes included in the training dataset along with the number of genomes and their source of availability.

C. difficile Ribotype	Number of Genomes	Source
RT001	24	Enterobase, NCBI, Creighton University
RT002	2	NCBI, Creighton University
RT003	19	NCBI, Creighton University
RT005	19	NCBI, Creighton University
RT010	3	NCBI, Creighton University
RT014	11	NCBI, Creighton University
RT015	2	NCBI, Creighton University
RT017	15	NCBI, Creighton University
RT023	3	NCBI, Creighton University
RT027	15	NCBI, Creighton University
RT046	4	NCBI, Creighton University
RT078	15	NCBI, Creighton University
RT106	22	Enterobase, NCBI, Creighton University
RT126	6	NCBI, Creighton University
TOTAL	**160**	

To identify associations between variant genetic loci and PCR RT, a hypothesis-free DBGWAS method was used. DBGWAS defines genetic variants linked to phenotypic traits via single nucleotide polymorphism (SNP), insertions, deletions, and consequences of recombination [41,42]. We used an open source tool (https://gitlab.com/leoisl/dbgwas) [40]. The tool is able to cover variants in coding as well as non-coding regions of bacterial genomes. DBGWAS was performed keeping the tool parameters in the default setting for different *C. difficile* ribotypes (RT001, RT017, RT027, RT078, and RT106) and their RT-specific genetic variants observed in the training set. Each *C. difficile* RT was considered independently in this study. DBGWAS identified short signature sequences called (overlapping) k-mers, yielding a compact summary of all variations across a set of genomes [40]. Overlapping k-mers are called unitigs and were selected on the basis of their specific and unique presence in a particular RT. Q-values define test sensitivity and specificity and are Benjamini–Hochberg-transformed *p*-values for controlling the false-positive results in case of multiple testing [40,43]. R scripting was used to deal with large matrices defining the presence (1) or absence (0) of extracted unitigs in the training set of *C. difficile* genomes.

2.3. Validation of Markers

Validation of novel unitig markers was performed by means of BLAST searches against the test set of genomes (Table 3). A wide range of 2425 genomes covering 132 different *C. difficile* RTs was downloaded [44] and processed using a Linux shell script. These sequences represented PCR ribotyped strains from different countries and clinical and environmental specimens for which phylogenetic analyses were already performed by Frentrup et al. [44] (Supplementary Table S3). A database of this test set was created to perform local command line BLAST searches against the set of significant unitigs identified above. The specificity of all the unitigs was tested using strict parametric filters of 100% coverage and identity.

Table 3. *C. difficile* ribotypes downloaded from the Enterobase database as a test dataset and the number of genomes included in each ribotype.

Ribotype	Count	Ribotype	Count	Ribotype	Count	Ribotype	Count
RT001	206	RT046	3	RT127	1	RT375	1
RT002	53	RT049	9	RT129	1	RT404	15
RT003	11	RT050	4	RT137	1	RT413	13
RT005	14	RT051	1	RT138	1	RT446	2
RT006	1	RT053	5	RT149	1	RT449	2
RT009	2	RT054	2	RT150	1	RT451	1
RT010	7	RT056	5	RT153	1	RT453	1
RT011	3	RT058	1	RT156	1	RT454	1
RT012	45	RT060	1	RT157	1	RT456	1
RT013	1	RT062	2	RT158	1	RT470	1
RT014	113	RT063	1	RT176	13	RT500	21
RT015	36	RT066	3	RT193	1	RT534	1
RT017	272	RT067	1	RT194	1	RT547	1
RT018	55	RT069	1	RT212	1	RT559	1
RT019	1	RT070	4	RT220	4	RT563	1
RT020	44	RT072	1	RT225	1	RT569	1
RT022	1	RT073	2	RT226	1	RT581	1
RT023	16	RT075	1	RT236	3	RT585	1
RT024	1	RT076	2	RT238	1	RT586	1
RT026	6	RT077	1	RT239	2	RT591	1
RT027	652	RT078	492	RT241	5	RT598	8
RT029	3	RT081	2	RT244	9	RT614	1
RT031	2	RT083	1	RT251	1	RT620	2
RT032	1	RT084	2	RT262	1	RT629	1
RT033	5	RT087	8	RT284	1	RT651	1
RT035	2	RT090	1	RT289	1	RT666	1
RT036	1	RT094	1	RT290	1	RT668	1
RT037	1	RT102	1	RT305	1	RT678	1
RT039	5	RT103	2	RT316	1	RT708	1
RT042	2	RT106	55	RT321	1	RT719	1
RT043	2	RT117	2	RT328	2	RT720	1
RT044	2	RT125	1	RT336	1	RT721	1
RT045	2	RT126	79	RT356	8	RT722	1

2.4. Statistically Reliable Ribotype Prediction

To evaluate the potential typing significance of the unitigs as compared to the classical ribotyping of *C. difficile* strains, sensitivity and specificity (selectivity) were computed for all the unitigs [45]. The efficiency of GWAS can be measured by assessing the false discovery rate (FDR) [46]. To increase the potential typing significance of our new method, combination statistics were performed. Sensitivity and specificity were also computed for certain combinations of two or more selected unitig sequences. The parameters defined were, next to the FDR, TP (true-positives, correctly predicting positive values, e.g., number of true RT017 predicted as RT017), FP (false-positives, missed negative values, e.g., number of non-RT017 genomes still predicted as RT017), FN (false-negatives, missed positive values), and TN (true-negatives, correctly rejected values).

2.5. Functional Annotation of Unitigs

Selected unitigs were annotated using BLASTn alignment. Well-characterized *C. difficile* genomes were used as a reference to locate these new markers. Specific annotation for each marker was filtered out using minimum E-value, 100% identity, 100% coverage, and 0 gap score.

3. Results and Discussion

3.1. In Silico PCR

The visualization of amplified DNA sequences from the intergenic region between 16S and 23S ribosomal genes is the current Gold Standard for *C. difficile* typing [25,47]. In our study, in silico PCR for 30 randomly selected, well-characterized *C. difficile* genome sequences was essentially unsuccessful (Figure 1 and Supplementary Table S1). Genome sequences included generated insufficient numbers of differently sized fragments. The fragment sizes that were calculated were verified with the online tool available at http://insilico.ehu.es/PCR [10]. On the other hand, more recently sequenced *C. difficile* genomes were showing only one or even none of the expected amplified fragments (Supplementary Table S1). There is a substantial possibility that the PCR ribotyping fragments observed upon laboratory experimentation for these strains may not derive from ISR variants but rather from random amplification. Thus, in silico PCR failed to generate reliable RTs which prompted us to explore the feasibility of DBGWAS-based typing. Of note, we presume here that NGS-based methods are very likely to be more reliable than any of the many other molecular typing methods.

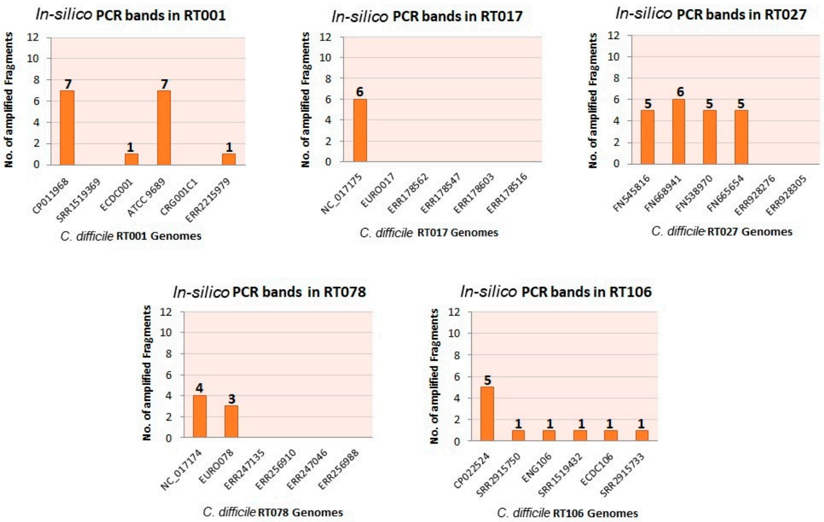


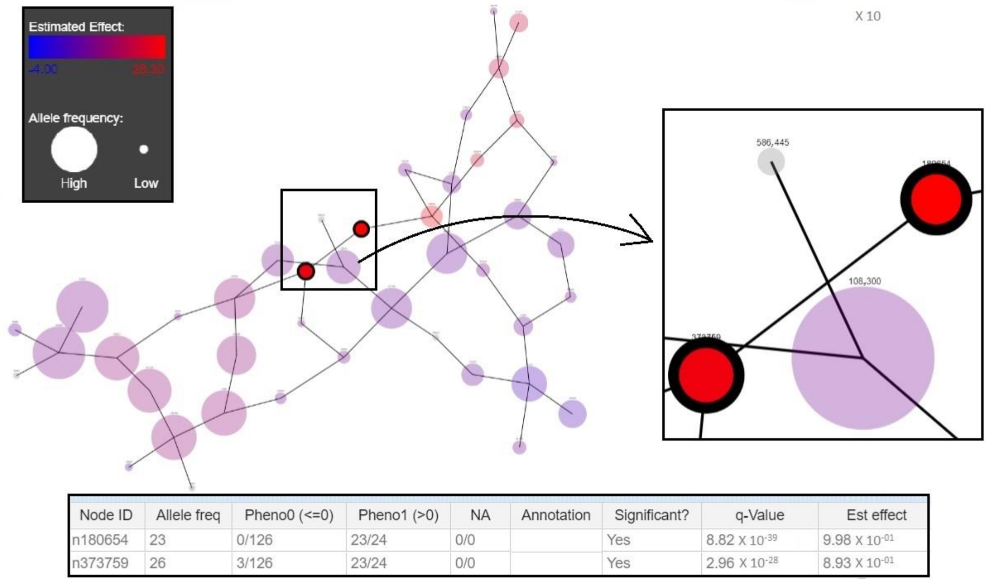
Figure 1. In silico ribotyping of different *C. difficile* genome sequences using the ISR 16S and 23S USA primer pair. The five panels represent the results obtained for examples of five different ribotypes. Bar graphs show the number of theoretical PCR bands (vertical axis, number of bands labeled on each bar) in the ribosomal region of respective genome sequences (horizontal axis), whereas the genomes without any fragments depict the complete absence of primer binding sites in those genomes. Note that the expected outcome would be an identical number of fragments for each of the strains belonging to a single ribotype. We indicated this number as the first marker of reproducibility; it has to be stated that besides this variation of numbers of fragments, the size of the fragment was also determined as a variable as well.

3.2. New Genotyping Markers

A total number of 47 RT-specific unitigs (13 for RT017, 12 for RT078, 9 for RT106, 7 for RT027, and 6 for RT001) were identified. The unitigs shared an average length of 56 base pairs (Table 4). DBGWAS generated compacted De Bruijn graphs (cDBG) containing the specific unitigs as nodes defining a genotypic association between a particular RT and the *C. difficile* genomes included (Figure 2). Unitigs that were specific for a particular RT were color-coded according to their association to the RT (red for positive association, blue for negative association) and minimum q-values were provided by subgraphs. Q-values for selected unitigs ranged from 0 to 7.40 × 10^{-5}. Significantly associated unitigs were extracted as FASTA formatted sequences (Supplementary Table S4).

Table 4. Number of unique markers identified for each *C. difficile* ribotype, their average length, and annotation.

Ribotype	Number of Markers	Average Length (Base Pairs)	Annotation (Number of Unitigs)
RT001	06	59	1. Intergenic (4) 2. tRNA uridine-5-carboxymethylaminomethyl synthesis enzyme MnmG (1) 3. rRNA-23S ribosomal RNA (1 excluded from the list) 4. Unknown (1)
RT017	13	69	1. Intergenic (3) 2. Membrane spanning protein (1) 3. Ribosome small subunit-dependent GTPase A (1) 4. Hypothetical protein (1) 5. EAL domain-containing protein (2) 6. Glutamate 2,3-aminomutase (1) 7. MurR/RpiR family transcriptional regulator (1) 8. GGDEF domain-containing protein (1) 9. Glyoxalase-like domain protein (1) 10. Radical SAM protein (1)
RT027	07	53	1. Collagen-like exosporium glycoprotein BclA2 (1) 2. Intergenic (5) 3. Unknown (1)
RT078	12	42	1. IS200/IS605 family element transposase accessory protein TnpB (1) 2. Spore surface glycoprotein BclB (4) 3. Collagen-like exosporium glycoprotein (BclA2) (3) 4. Unknown (partial with ABC transporter permease) (1) 5. Intergenic (1) 6. Site-specific integrase (1) 7. S8 family peptidase (1)
RT106	09	55	1. Intergenic (3) 2. ABC transporter permease (2) 3. Hypothetical Protein (1) 4. 3-Hydroxybutyryl-CoA dehydrogenase (1) 5. Potassium transporter (2)



Node ID	Allele freq	Pheno0 (<=0)	Pheno1 (>0)	NA	Annotation	Significant?	q-Value	Est effect
n180654	23	0/126	23/24	0/0		Yes	8.82 X 10^{-39}	9.98 X 10^{-01}
n373759	26	3/126	23/24	0/0		Yes	2.96 X 10^{-28}	8.93 X 10^{-01}

Figure 2. Compacted De Bruijn graphs (cDBG) generated by De Bruijn graph-based Genome Wide Association Studies (DBGWAS) for RT001 genome sequences. The figure illustrates the significance of the nodes (representing the selective sequences called unitigs), which are denoted by their estimated effect ranging from high (28.304; red) to low (4.00; blue). Allele frequency is represented by the size of the node. The table explains that from the two selected significant nodes in terms of their association with ribotype, the node on the top right (n180654) is specific to RT001 (called Pheno 1 in the table) and completely absent in the other ribotypes in the training set (Pheno 0). Additionally, the q-value linked to the first node is very significantly below 0.05 and hence, the estimated effect is high (represented by the red color of the node).

3.3. Validation of Markers

Unitigs showing 100% identity in all genomes belonging to a single RT in the validation set demonstrated the efficiency of these unique patterns to carry out in silico ribotyping. Although the individual unitig-based characterization of *C. difficile* strains was not absolute, it allowed RT determination with approximate sensitivity and specificity of between 0.90 and 1.0 (Figure 3). FDR for all the unitigs for RT017 was the lowest (0.06) followed by RT027 (0.08), RT078 (0.23), RT001 (0.30), and RT106 (0.46) (Figure 3).

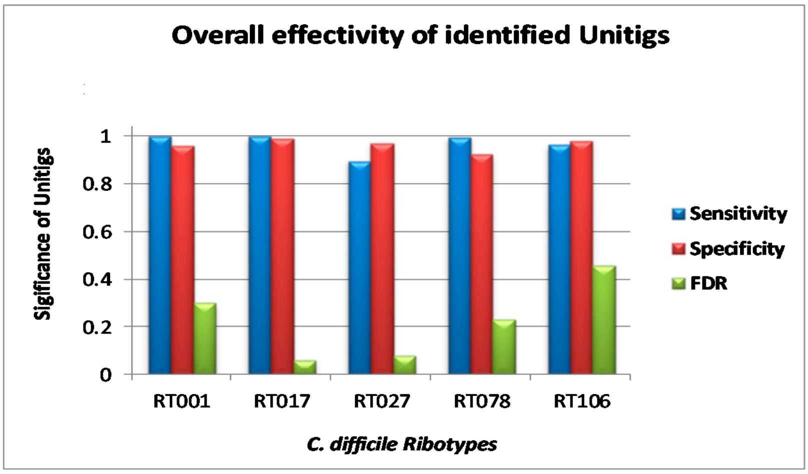
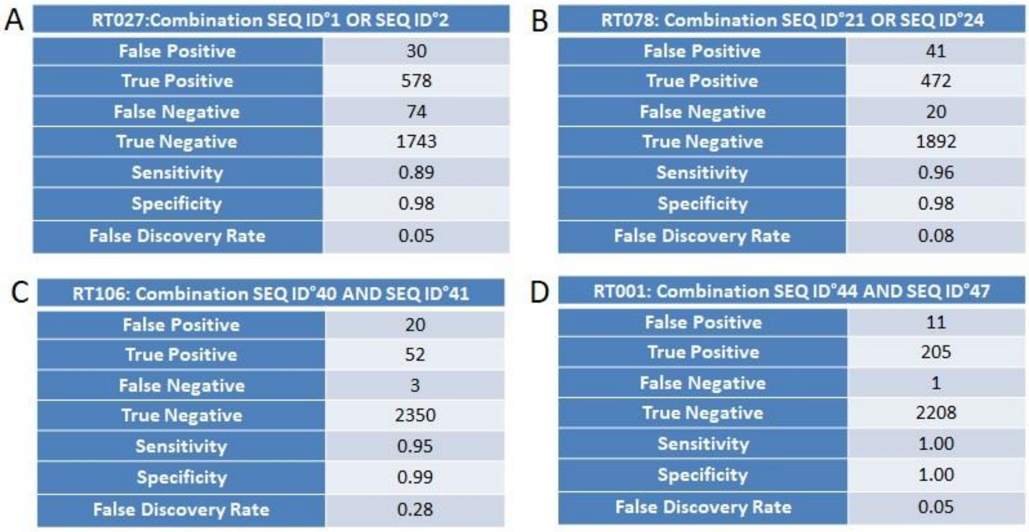


Figure 3. Statistical comparison of genome typing efficiency of discovered unique patterns for selected *C. difficile* ribotypes in terms of specificity, sensitivity, and false discovery rate (FDR).

Some of the unitigs were shared by closely related RTs. Unitigs for RT001 were able to identify the genetically closely related RT087, RT241, and RT012, which altogether form a clonal complex (CC) 141 [44]. One of the markers identified for RT017 showed no false-positives or false negatives. Other markers for RT017 initially generated a small number of false-positives, but 100% true-positives in the validation dataset. Markers for RT078 identified 78 out of 79 isolates of RT126 and all of the RT413 strains from the test dataset, likely due to the close genetic relatedness of these RTs (CC 1) [44,48,49]. Unitig sequences for RT106 were also able to identify *C. difficile* RT500 along with RT106 from the test set. Phylogenetic grouping of *C. difficile* genomes [44] showed that *C. difficile* core genome multi locus sequence typing (cgMLST) of RT106 and RT500 (CC 22) generated completely indistinguishable groupings. Considering closely related strains as true-positives based on their respective RTs, the FDRs for each unitig subset were found to be smaller, underscoring the biological consistency of the results. Adding genomes of RT413, RT126, and RT500 to the training set resulted in a decreased FDR rate. The continuously increasing number of publicly available *C. difficile* genome sequences will provide substantial opportunities for improvement of our new characterization technique.

3.4. Marker Combination Study

For the ribotypes RT027, RT078, RT106, and RT001, every possible combination of RT-specific unitigs was created and tested for statistical significance. A combination of two unitigs was shown to increase sensitivity and specificity up to 1 and to reduce the FDR to 0.05 (Figure 4A–D). Each combination was defined on the basis of logical operators "AND/OR". The AND operator symbolizes that both the markers in a combination need to be present with 100% identity, whereas the OR operator means that any one of the two markers in a combination need to be present at one time, again with 100% sequence identity. There is no combination required in the case of RT017. Conclusively, as clearly exemplified in Figure 4A–D, in certain cases, the combination of markers improves RT testing by suppression of the false discovery rate. Marker's SEQ ID numbers and their sequences are given in Supplementary Table S4.



A

RT027:Combination SEQ ID°1 OR SEQ ID°2	
False Positive	30
True Positive	578
False Negative	74
True Negative	1743
Sensitivity	0.89
Specificity	0.98
False Discovery Rate	0.05

B

RT078: Combination SEQ ID°21 OR SEQ ID°24	
False Positive	41
True Positive	472
False Negative	20
True Negative	1892
Sensitivity	0.96
Specificity	0.98
False Discovery Rate	0.08

C

RT106: Combination SEQ ID°40 AND SEQ ID°41	
False Positive	20
True Positive	52
False Negative	3
True Negative	2350
Sensitivity	0.95
Specificity	0.99
False Discovery Rate	0.28

D

RT001: Combination SEQ ID°44 AND SEQ ID°47	
False Positive	11
True Positive	205
False Negative	1
True Negative	2208
Sensitivity	1.00
Specificity	1.00
False Discovery Rate	0.05

Figure 4. (**A**–**D**) Statistical reliability in terms of sensitivity, specificity, and false discovery rate (FDR) for the combination of two selected markers using OR operator for the identification of *C. difficile* RT027 (Panel **A**) and RT078 (Panel **B**). Panels **C** and **D** display similar analyses but then using the AND operator for identification of RT106 and RT001, respectively.

3.5. Functional Annotation of Markers

Functional characterization of the regions from which our unitigs originated demonstrated that 34% of the unitigs were localized in intergenic regions (five for RT027, four for RT001, three for of RT017 and RT106 each, and one for RT078 (Figures 5–9). Six percent of all markers were left unannotated in

RT001, RT027, and RT078 (one marker for each) (Table 4). Only RT001 was identified with a unitig marker residing within the rRNA-23S ribosomal gene showing at least some correspondence with ribotyping (Figure 5). This marker did not show sufficient diagnostic power and was thus not selected in the final set of markers. All other markers were observed to be scattered throughout the *C. difficile* genome. In RT078, one of these markers was identified in a mobile genetic element (Figure 8). Mostly, genes and intergenic regions, apart from the conserved ribosomal ISR, were observed to play a potential role in the unitig-mediated *C. difficile* typing.

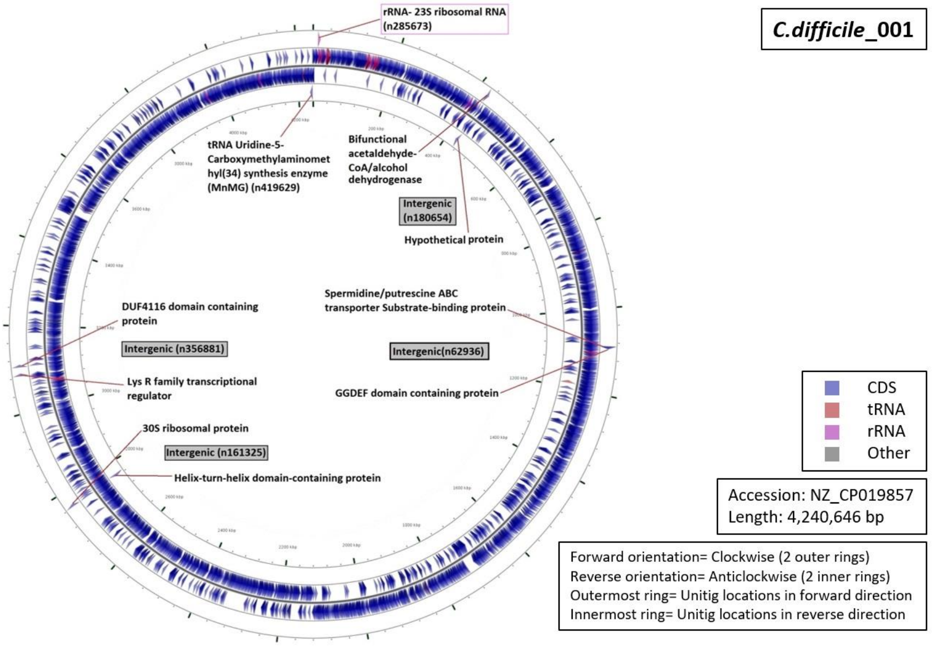

rRNA- 23S ribosomal RNA (n285673)
C.difficile_001
tRNA Uridine-5-Carboxymethylaminomethyl(34) synthesis enzyme (MnMG) (n419629)
Bifunctional acetaldehyde-CoA/alcohol dehydrogenase
Intergenic (n180654)
Hypothetical protein
Spermidine/putrescine ABC transporter Substrate-binding protein
DUF4116 domain containing protein
Intergenic (n356881)
Intergenic(n62936)
Lys R family transcriptional regulator
GGDEF domain containing protein
30S ribosomal protein
Intergenic (n161325)
Helix-turn-helix domain-containing protein
CDS
tRNA
rRNA
Other
Accession: NZ_CP019857
Length: 4,240,646 bp
Forward orientation= Clockwise (2 outer rings)
Reverse orientation= Anticlockwise (2 inner rings)
Outermost ring= Unitig locations in forward direction
Innermost ring= Unitig locations in reverse direction

Figure 5. Functional annotation and location of DBGWAS markers on the reference genome of *C. difficile* RT001. Both central rings represent the genome annotation (reverse inside, forward outside), while the outer and inner rings represent the signature sequences (unitigs) (reverse inside, forward outside).

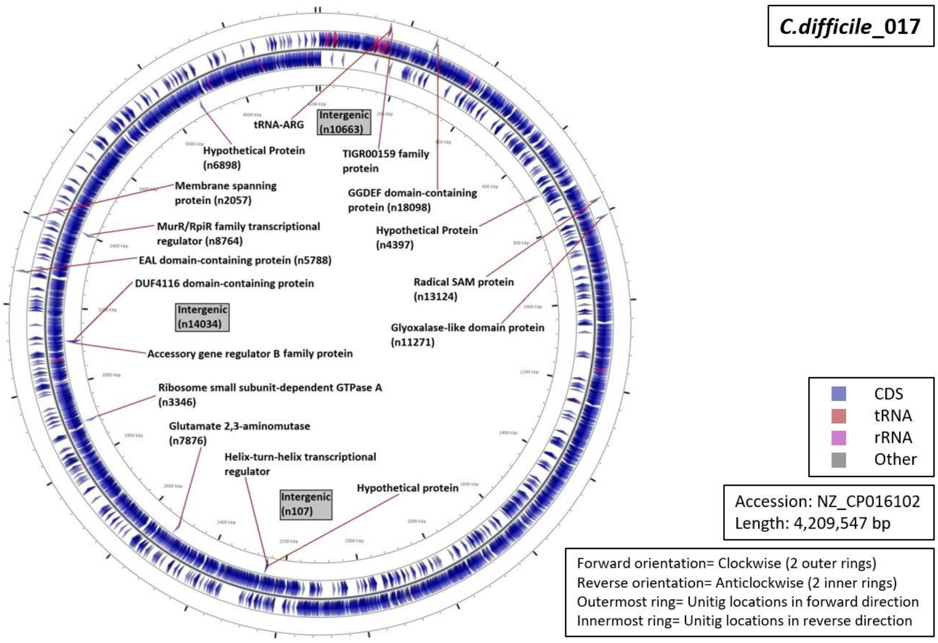

C.difficile_017
tRNA-ARG
Intergenic (n10663)
Hypothetical Protein (n6898)
TIGR00159 family protein
Membrane spanning protein (n2057)
GGDEF domain-containing protein (n18098)
MurR/RpiR family transcriptional regulator (n8764)
Hypothetical Protein (n4397)
EAL domain-containing protein (n5788)
DUF4116 domain-containing protein
Radical SAM protein (n13124)
Intergenic (n14034)
Glyoxalase-like domain protein (n11271)
Accessory gene regulator B family protein
Ribosome small subunit-dependent GTPase A (n3346)
Glutamate 2,3-aminomutase (n7876)
Helix-turn-helix transcriptional regulator
Intergenic (n107)
Hypothetical protein
CDS
tRNA
rRNA
Other
Accession: NZ_CP016102
Length: 4,209,547 bp
Forward orientation= Clockwise (2 outer rings)
Reverse orientation= Anticlockwise (2 inner rings)
Outermost ring= Unitig locations in forward direction
Innermost ring= Unitig locations in reverse direction

Figure 6. Functional annotation and location of DBGWAS markers on the reference genome of *C. difficile* RT017.

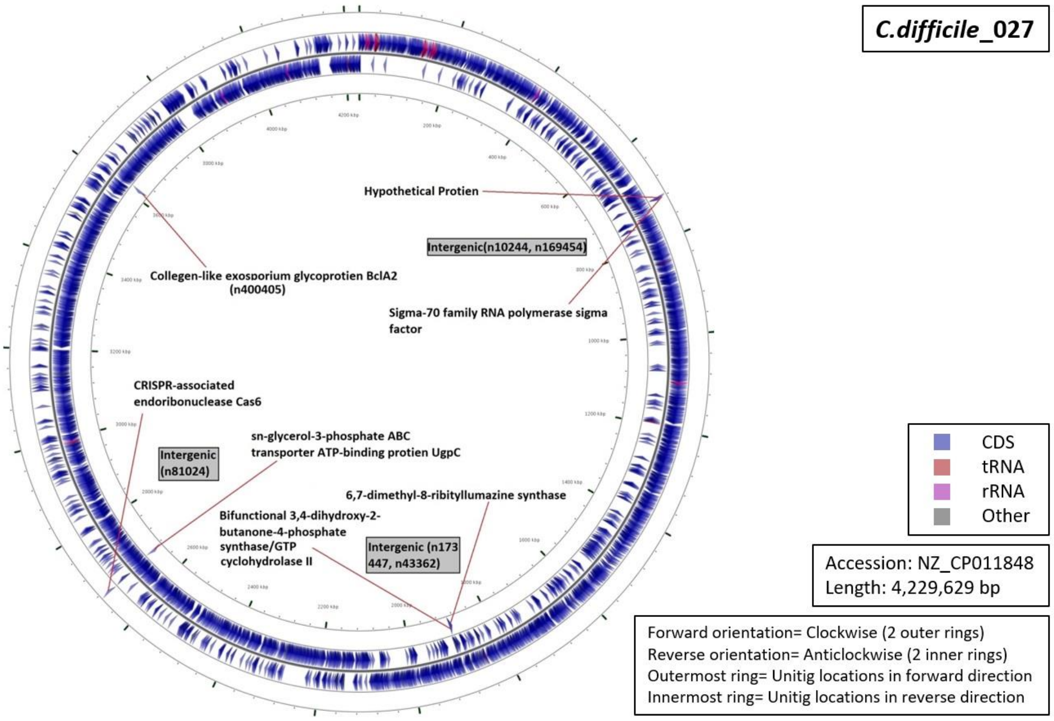

C.difficile_027
Hypothetical Protien
Intergenic(n10244, n169454)
Collegen-like exosporium glycoprotien BclA2 (n400405)
Sigma-70 family RNA polymerase sigma factor
CRISPR-associated endoribonuclease Cas6
sn-glycerol-3-phosphate ABC transporter ATP-binding protien UgpC
Intergenic (n81024)
6,7-dimethyl-8-ribityllumazine synthase
Bifunctional 3,4-dihydroxy-2-butanone-4-phosphate synthase/GTP cyclohydrolase II
Intergenic (n173 447, n43362)
CDS
tRNA
rRNA
Other
Accession: NZ_CP011848
Length: 4,229,629 bp
Forward orientation= Clockwise (2 outer rings)
Reverse orientation= Anticlockwise (2 inner rings)
Outermost ring= Unitig locations in forward direction
Innermost ring= Unitig locations in reverse direction

Figure 7. Functional annotation and location of DBGWAS markers on the reference genome of *C. difficile* RT027.

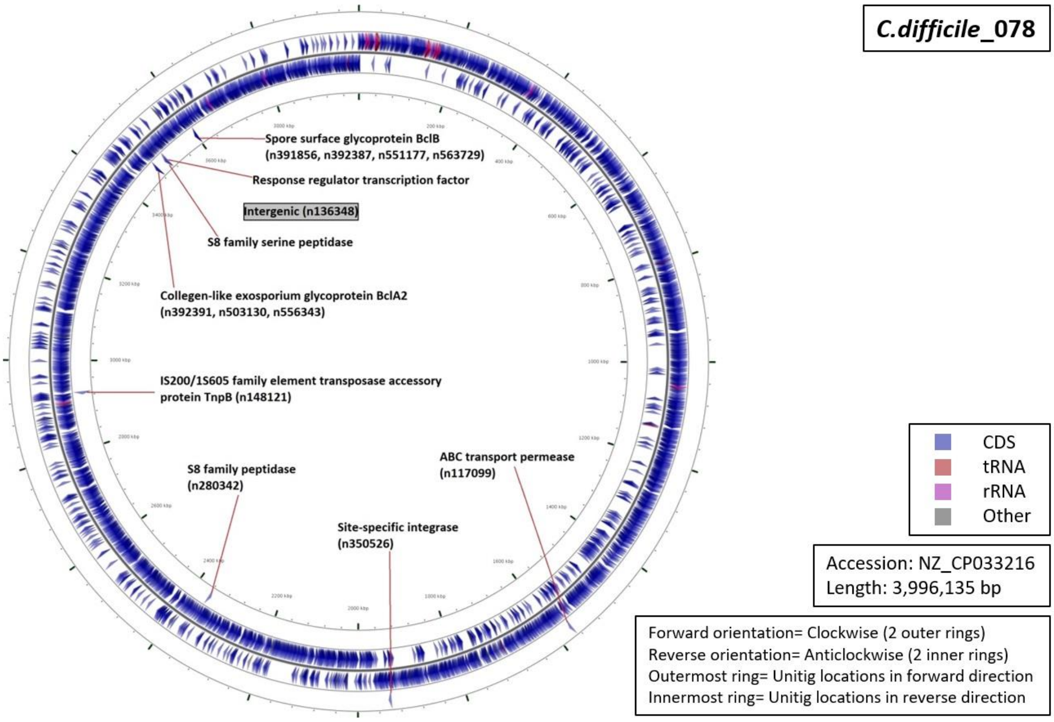

C.difficile_078
Spore surface glycoprotein BclB (n391856, n392387, n551177, n563729)
Response regulator transcription factor
Intergenic (n136348)
S8 family serine peptidase
Collegen-like exosporium glycoprotein BclA2 (n392391, n503130, n556343)
IS200/1S605 family element transposase accessory protein TnpB (n148121)
S8 family peptidase (n280342)
ABC transport permease (n117099)
Site-specific integrase (n350526)
CDS
tRNA
rRNA
Other
Accession: NZ_CP033216
Length: 3,996,135 bp
Forward orientation= Clockwise (2 outer rings)
Reverse orientation= Anticlockwise (2 inner rings)
Outermost ring= Unitig locations in forward direction
Innermost ring= Unitig locations in reverse direction

Figure 8. Functional annotation and location of DBGWAS markers on the reference genome of *C. difficile* RT078.

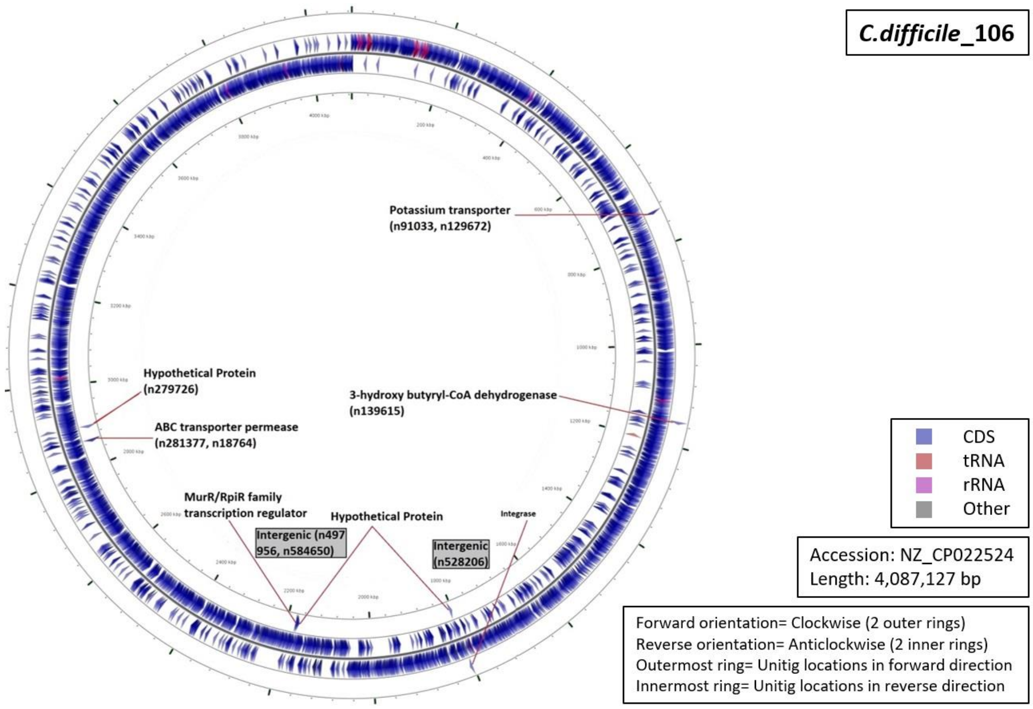

C.difficile_106
Potassium transporter (n91033, n129672)
Hypothetical Protein (n279726)
3-hydroxy butyryl-CoA dehydrogenase (n139615)
ABC transporter permease (n281377, n18764)
MurR/RpiR family transcription regulator
Hypothetical Protein
Integrase
Intergenic (n497956, n584650)
Intergenic (n528206)
CDS
tRNA
rRNA
Other
Accession: NZ_CP022524
Length: 4,087,127 bp
Forward orientation= Clockwise (2 outer rings)
Reverse orientation= Anticlockwise (2 inner rings)
Outermost ring= Unitig locations in forward direction
Innermost ring= Unitig locations in reverse direction

Figure 9. Functional annotation and location of DBGWAS markers on the reference genome of *C. difficile* RT106.

4. Conclusions

Strain typing has a proven value in monitoring the persistence and spread of bacterial pathogens in human populations. For *C. difficile*, PCR ribotyping is the current first choice but may be challenged now that genome sequencing is an option. No single-step test or algorithm is available so far for correlating *C. difficile* RTs with WGS data. This implies that there may be an issue with the correlation between WGS-based epidemiological analysis and PCR ribotyping for *C. difficile*. Here, we show that DBGWAS identified unique genomic markers that would suit that specific purpose. A combination of two unitigs led to 100% sensitive and specific discrimination between five important RTs. We believe that this approach is highly promising, providing a clear opportunity to define backward compatibility between classical RTs and WGS data.

Supplementary Materials: The following are available online at http://www.mdpi.com/2075-4418/10/12/1078/s1. Table S1 describes a collection of genome sequences that were used for the in silico search of ribotypes based on amplification of tentative ISRs. Table S2 contains a training set of *C. difficile* genomes used for the initial DBGWAS. Table S3 contains the *C. difficile* genomes used for DBGWAS validation. Table S4 shows a review of all unitigs that are statistically significantly associated with specific ribotypes.

Author Contributions: Conceptualization, A.v.B., M.G., L.H., M.J., K.D.B.and R.V.G.; Formal analysis, M.G.; Investigation, M.G., A.v.B., L.H., M.J. and R.V.G.; Methodology, M.G. and K.D.B.; Software, M.J.; Supervision, A.v.B.; Writing—original draft, M.G.; Writing— review & editing, A.v.B., L.H., H.P., M.J., K.D.B.and R.G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported and funded by bioMérieux, France; Creighton University School of Medicine, NE, USA; and the European Union's Horizon 2020 research and innovation program entitled as Viral and Bacterial Adhesin Network Training (ViBrANT) under Marie Skłodowska-Curie Grant Agreement No. 765042.

Acknowledgments: During this study, M.G., L.H., H.P., M.J., K.D.B. and A.v.B. were employees of bioMérieux, a company designing, developing, and marketing tests in the domain of infectious diseases. The company was not involved in the design of the current study and the opinions expressed are those of the authors and may be different from formal company opinions and policies. We thank our colleagues from bioMérieux and Creighton University who provided expertise and insight that greatly assisted the research.

Conflicts of Interest: This research was conducted in association with RG from Creighton University School of Medicine, NE, USA which could be constructed as a potential conflict of interest.

References

1. Walk, S.T.; Micic, D.; Jain, R.; Lo, E.S.; Trivedi, I.; Liu, E.W.; Almassalha, L.M.; Ewing, S.A.; Ring, C.; Galecki, A.T.; et al. *Clostridium difficile* Ribotype Does Not Predict Severe Infection. *Clin. Infect. Dis.* **2012**, *55*, 1661–1668. [CrossRef] [PubMed]
2. Walker, A.S.; Eyre, D.W.; Wyllie, D.H.; Dingle, K.E.; Griffiths, D.; Shine, B.; Walker, A.S.; O'Connor, L.; Finney, J.; Vaughan, A.; et al. Relationship Between Bacterial Strain Type, Host Biomarkers, and Mortality in *Clostridium difficile* Infection. *Clin. Infect. Dis.* **2013**, *56*, 1589–1600. [CrossRef] [PubMed]
3. McFarland, L.V. Renewed interest in a difficult disease: *Clostridium difficile* infections—Epidemiology and current treatment strategies. *Curr. Opin. Gastroenterol.* **2009**, *25*, 24–35. [CrossRef] [PubMed]
4. Lessa, F.C.; Gould, C.V.; McDonald, L.C. Current Status of *Clostridium difficile* Infection Epidemiology. *Clin. Infect. Dis.* **2012**, *55*, S65–S70. [CrossRef] [PubMed]
5. Wiegand, P.N.; Nathwani, D.; Wilcox, M.H.; Stephens, J.; Shelbaya, A.; Haider, S. Clinical and economic burden of *Clostridium difficile* infection in Europe: A systematic review of healthcare-facility-acquired infection. *J. Hosp. Infect.* **2012**, *81*, 1–14. [CrossRef] [PubMed]
6. Deshpande, A.; Pasupuleti, V.; Thota, P.; Pant, C.; Rolston, D.D.K.; Sferra, T.J.; Hernandez, A.V.; Donskey, C.J. Community-associated *Clostridium difficile* infection and antibiotics: A meta-analysis. *J. Antimicrob. Chemother.* **2013**, *68*, 1951–1961. [CrossRef] [PubMed]
7. Namiki, H.; Kobayashi, T. Long-term, low-dose of clarithromycin as a cause of community-acquired *Clostridium difficile* infection in a 5-year-old boy. *Oxf. Med. Case Rep.* **2018**, *2018*, omx106. [CrossRef]
8. Lessa, F.C.; Mu, Y.; Bamberg, W.M.; Beldavs, Z.G.; Dumyati, G.K.; Dunn, J.R.; Farley, M.M.; Holzbauer, S.M.; Meek, J.I.; Phipps, E.C.; et al. Burden of *Clostridium difficile* infection in the United States. *N. Engl. J. Med.* **2015**, *372*, 825–834. [CrossRef]
9. Davies, K.G.; Longshaw, C.M.; Davis, G.L.; Bouza, E.; Barbut, F.; Barna, Z.; Delmée, M.; Fitzpatrick, F.; Ivanova, K.; Kuijper, E.; et al. Underdiagnosis of *Clostridium difficile* across Europe: The European, multicentre, prospective, biannual, point-prevalence study of *Clostridium difficile* infection in hospitalised patients with diarrhoea (EUCLID). *Lancet Infect. Dis.* **2014**, *14*, 1208–1219. [CrossRef]
10. Borren, N.Z.; Ghadermarzi, S.; Hutfless, S.; Ananthakrishnan, A.N. The emergence of *Clostridium difficile* infection in Asia: A systematic review and meta-analysis of incidence and impact. *PLoS ONE* **2017**, *12*, e0176797. [CrossRef]
11. Balsells, E.; Shi, T.; Leese, C.; Lyell, I.; Burrows, J.; Wiuff, C.; Campbell, H.; Kyaw, M.H.; Nair, H. Global burden of *Clostridium difficile* infections: A systematic review and meta-analysis. *J. Glob. Health* **2018**, *9*, 010407. [CrossRef] [PubMed]
12. Mills, J.P.; Rao, K.; Young, V.B. Probiotics for prevention of *Clostridium difficile* infection. *Curr. Opin. Gastroenterol.* **2018**, *34*, 3–10. [CrossRef] [PubMed]
13. CDC. *Biggest Threats Antibiotic/Antimicrobial Resistance*; CDC: Atlanta, GA, USA, 2017.
14. Burnham, C.-A.D.; Carroll, K.C. Diagnosis of *Clostridium difficile* Infection: An Ongoing Conundrum for Clinicians and for Clinical Laboratories. *Clin. Microbiol. Rev.* **2013**, *26*, 604–630. [CrossRef] [PubMed]
15. Planche, T.; Wilcox, M. Reference assays for *Clostridium difficile* infection: One or two gold standards? *J. Clin. Pathol.* **2011**, *64*, 1–5. [CrossRef]
16. She, R.C.; Durrant, R.J.; Petti, C.A. Evaluation of Enzyme Immunoassays to Detect *Clostridium difficile* Toxin from Anaerobic Stool Culture. *Am. J. Clin. Pathol.* **2009**, *131*, 81–84. [CrossRef]
17. Shetty, N.; Wren, M.; Coen, P. The role of glutamate dehydrogenase for the detection of *Clostridium difficile* in faecal samples: A meta-analysis. *J. Hosp. Infect.* **2011**, *77*, 1–6. [CrossRef]
18. Eckert, C.; Jones, G.; Barbut, F. Diagnosis of *Clostridium difficile* infection: The molecular approach. *Future Microbiol.* **2013**, *8*, 1587–1598. [CrossRef]
19. Krutova, M.; Wilcox, M.; Kuijper, E. A two-step approach for the investigation of a *Clostridium difficile* outbreak by molecular methods. *Clin. Microbiol. Infect.* **2019**, *25*, 1300–1301. [CrossRef]
20. Chatterjee, S.; Raval, I.H. Chapter 32—Pathogenic Microbial Genetic Diversity with Reference to Health. In *Microbial Diversity in the Genomic Era*; Academic Press: Cambridge, MA, USA, 2019; pp. 559–577.
21. Dingle, T.C.; MacCannell, D.R. Chapter 9—Molecular Strain Typing and Characterisation of Toxigenic *Clostridium difficile*. In *Methods in Microbiology*; Elsevier: Amsterdam, The Netherlands, 2015; pp. 329–357.

22. Krutova, M.; Kinross, P.; Barbut, F.; Hajdu, A.; Wilcox, M.; Kuijper, E.; Allerberger, F.; Delmée, M.; Van Broeck, J.; Vatcheva-Dobrevska, R.; et al. How to: Surveillance of *Clostridium difficile* infections. *Clin. Microbiol. Infect.* **2018**, *24*, 469–475. [CrossRef]
23. Collins, D.A.; Elliott, B.; Riley, T.V. Molecular methods for detecting and typing of *Clostridium difficile*. *Pathology* **2015**, *47*, 211–218. [CrossRef]
24. Bidet, P.; Barbut, F.; Lalande, V.; Burghoffer, B.; Petit, J.-C. Development of a new PCR-ribotyping method for *Clostridium difficile* based on ribosomal RNA gene sequencing. *FEMS Microbiol. Lett.* **1999**, *175*, 261–266. [CrossRef] [PubMed]
25. Indra, A.; Blaschitz, M.; Kernbichler, S.; Reischl, U.; Wewalka, G.; Allerberger, F. Mechanisms behind variation in the *Clostridium difficile* 16S–23S rRNA intergenic spacer region. *J. Med. Microbiol.* **2010**, *59*, 1317–1323. [CrossRef] [PubMed]
26. Indra, A.; Huhulescu, S.; Schneeweis, M.; Hasenberger, P.; Kernbichler, S.; Fiedler, A.; Wewalka, G.; Allerberger, F.; Kuijper, E.J. Characterization of *Clostridium difficile* isolates using capillary gel electrophoresis-based PCR ribotyping. *J. Med. Microbiol.* **2008**, *57*, 1377–1382. [CrossRef] [PubMed]
27. Waslawski, S.; Lo, E.S.; Ewing, S.A.; Young, V.B.; Aronoff, D.M.; Sharp, S.E.; Novak-Weekley, S.M.; Crist, A.E.; Dunne, W.M.; Hoppe-Bauer, J.; et al. *Clostridium difficile* Ribotype Diversity at Six Health Care Institutions in the United States. *J. Clin. Microbiol.* **2013**, *51*, 1938–1941. [CrossRef] [PubMed]
28. Janezic, S. Direct PCR-Ribotyping of *Clostridium difficile*. In *Clostridium difficile Methods in Molecular Biology*; Humana Press: New York, NY, USA, 2016; Volume 1476.
29. Martinez-Melendez, A.; Camacho-Ortiz, A.; Morfin-Otero, R.; Maldonado-Garza, H.J.; Villarreal-Trevino, L.; Garza-Gonzalez, E. Current knowledge on the laboratory diagnosis of *Clostridium difficile* infection. *World J. Gastroenterol.* **2017**, *23*, 1552–1567. [CrossRef] [PubMed]
30. Polage, C.R.; Gyorke, C.E.; Kennedy, M.A.; Leslie, J.L.; Chin, D.L.; Wang, S.; Nguyen, H.H.; Huang, B.; Tang, Y.W.; Lee, L.W.; et al. Overdiagnosis of *Clostridium difficile* Infection in the Molecular Test Era. *JAMA Intern. Med.* **2015**, *175*, 1792–1801. [CrossRef]
31. Janezic, S.; Rupnik, M. Development and Implementation of Whole Genome Sequencing-Based Typing Schemes for *Clostridioides difficile*. *Front. Public Health* **2019**, *7*. [CrossRef]
32. Fawley, W.N.; Knetsch, C.W.; MacCannell, D.R.; Harmanus, C.; Du, T.; Mulvey, M.R.; Paulick, A.; Anderson, L.; Kuijper, E.J.; Wilcox, M.H. Development and Validation of an Internationally-Standardized, High-Resolution Capillary Gel-Based Electrophoresis PCR-Ribotyping Protocol for *Clostridium difficile*. *PLoS ONE* **2015**, *10*, e0118150. [CrossRef]
33. Bauer, M.P.; Notermans, D.W.; Van Benthem, B.H.; Brazier, J.S.; Wilcox, M.H.; Rupnik, M.; Monnet, D.L.; Van Dissel, J.T.; Kuijper, E.J. *Clostridium difficile* infection in Europe: A hospital-based survey. *Lancet Infect. Dis.* **2011**, *377*, 63–73. [CrossRef]
34. Arvand, M.; Hauri, A.M.; Zaiss, N.H.; Witte, W.; Bettge-Weller, G. *Clostridium difficile* ribotypes 001, 017, and 027 are associated with lethal *C. difficile* infection in Hesse, Germany. *Eurosurveillance* **2009**, *14*, 19403. [CrossRef]
35. Vanek, J.; Hill, K.; Collins, J.; Berrington, A.; Perry, J.; Inns, T.; Gorton, R.; Magee, J.; Sails, A.; Mullan, A.; et al. Epidemiological survey of *Clostridium difficile* ribotypes in the North East of England during an 18-month period. *J. Hosp. Infect.* **2012**, *81*, 209–212. [CrossRef] [PubMed]
36. Giancola, S.E.; Williams, R.J.; Gentry, C.A. Prevalence of the *Clostridium difficile* BI/NAP1/027 strain across the United States Veterans Health Administration. *Clin. Microbiol. Infect.* **2018**, *24*, 877–881. [CrossRef] [PubMed]
37. Howell, M.D.; Novack, V.; Grgurich, P.; Soulliard, D.; Novack, L.; Pencina, M.; Talmor, D. Latrogenic gastric acid suppression and the risk of nosocomial *Clostridium difficile* infection. *Arch. Intern. Med.* **2010**, *170*, 784–790. [CrossRef] [PubMed]
38. Imwattana, K.; Knight, D.R.; Kullin, B.; Collins, D.A.; Putsathit, P.; Kiratisin, P.; Riley, T.V. *Clostridium difficile* ribotype 017—Characterization, evolution and epidemiology of the dominant strain in Asia. *Emerg. Microbes Infect.* **2019**, *8*, 796–807. [CrossRef] [PubMed]
39. Kim, J.; Kim, Y.; Pai, H. Clinical Characteristics and Treatment Outcomes of *Clostridium difficile* Infections by PCR Ribotype 017 and 018 Strains. *PLoS ONE* **2016**, *11*, e0168849. [CrossRef]

40. Jaillard, M.; Lima, L.; Tournoud, M.; Mahé, P.; Van Belkum, A.; Lacroix, V.; Jacob, L. A fast and agnostic method for bacterial genome-wide association studies: Bridging the gap between k-mers and genetic events. *PLoS Genet.* **2018**, *14*, e1007758. [CrossRef]
41. Alam, M.T.; Petit, R.A.; Crispell, E.K.; Thornton, T.A.; Conneely, K.N.; Jiang, Y.; Satola, S.W.; Read, T.D. Dissecting vancomycin-intermediate resistance in staphylococcus aureus using genome-wide association. *Genome Biol. Evol.* **2014**, *6*, 1174–1185. [CrossRef]
42. Chewapreecha, C.; Marttinen, P.; Croucher, N.J.; Salter, S.J.; Harris, S.R.; Mather, A.E.; Hanage, W.P.; Goldblatt, D.; Nosten, F.H.; Turner, C.; et al. Comprehensive Identification of Single Nucleotide Polymorphisms Associated with Beta-lactam Resistance within Pneumococcal Mosaic Genes. *PLoS Genet.* **2014**, *10*, e1004547. [CrossRef]
43. Benjamini, Y.; Hochberg, Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **1995**, *57*, 289–300. [CrossRef]
44. Frentrup, M.; Zhou, Z.; Steglich, M.; Meier-Kolthoff, J.P.; Göker, M.; Riedel, T.; Bunk, B.; Spröer, C.; Overmann, J.; Blaschitz, M.; et al. A publicly accessible database for *Clostridioides difficile* genome sequences supports tracing of transmission chains and epidemics. *Microb. Genom.* **2020**, *6*, 410. [CrossRef]
45. Baratloo, A.; Hosseini, M.; Negida, A.; El Ashal, G. Simple Definition and Calculation of Accuracy, Sensitivity and Specificity. *Emergency* **2015**, *3*, 48–49. [PubMed]
46. Bradbury, P.; Parker, T.; Hamblin, M.T.; Jannink, J. Assessment of Power and False Discovery Rate in Genome-Wide Association Studies using the BarleyCAP Germplasm. *Crop. Sci.* **2011**, *51*, 52–59. [CrossRef]
47. Xiao, M.; Kong, F.; Jin, P.; Wang, Q.; Xiao, K.; Jeoffreys, N.; James, G.; Gilbert, G.L. Comparison of Two Capillary Gel Electrophoresis Systems for *Clostridium difficile* Ribotyping, Using a Panel of Ribotype 027 Isolates and Whole-Genome Sequences as a Reference Standard. *J. Clin. Microbiol.* **2012**, *50*, 2755–2760. [CrossRef] [PubMed]
48. Alvarez-Perez, S.; Blanco, J.L.; Harmanus, C.; Kuijper, E.; Garcia, M.E. Subtyping and antimicrobial susceptibility of *Clostridium difficile* PCR ribotype 078/126 isolates of human and animal origin. *Vet. Microbiol.* **2017**, *199*, 15–22. [CrossRef]
49. Schneeberg, A.; Neubauer, H.; Schmoock, G.; Baier, S.; Harlizius, J.; Nienhoff, H.; Brase, K.; Zimmermann, S.; Seyboldt, C. *Clostridium difficile* Genotypes in Piglet Populations in Germany. *J. Clin. Microbiol.* **2013**, *51*, 3796–3803. [CrossRef]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.