scientific reports

Check for updates

OPEN

Phylogenomic analysis of *Clostridioides difficile* ribotype 106 strains reveals novel genetic islands and emergent phenotypes

Bryan Angelo P. Roxas[1,7], Jennifer Lising Roxas[1,7], Rachel Claus-Walker[1], Anusha Harishankar[1], Asad Mansoor[1], Farhan Anwar[1], Shobitha Jillella[1], Alison Williams[1], Jason Lindsey[1], Sean P. Elliott[2], Kareem W. Shehab[2], V. K. Viswanathan[1,3,4] & Gayatri Vedantam[1,3,4,5,6]✉

Clostridioides difficile **infection (CDI) is a major healthcare-associated diarrheal disease. Consistent with trends across the United States, *C. difficile* RT106 was the second-most prevalent molecular type in our surveillance in Arizona from 2015 to 2018. A representative RT106 strain displayed robust virulence and 100% lethality in the hamster model of acute CDI. We identified a unique 46 KB genomic island (GI1) in all RT106 strains sequenced to date, including those in public databases. GI1 was not found in its entirety in any other *C. difficile* clade, or indeed, in any other microbial genome; however, smaller segments were detected in *Enterococcus faecium* strains. Molecular clock analyses suggested that GI1 was horizontally acquired and sequentially assembled over time. GI1 encodes homologs of VanZ and a SrtB-anchored collagen-binding adhesin, and correspondingly, all tested RT106 strains had increased teicoplanin resistance, and a majority displayed collagen-dependent biofilm formation. Two additional genomic islands (GI2 and GI3) were also present in a subset of RT106 strains. All three islands are predicted to encode mobile genetic elements as well as virulence factors. Emergent phenotypes associated with these genetic islands may have contributed to the relatively rapid expansion of RT106 in US healthcare and community settings.**

Abbreviations

CDI	*Clostridioides difficile* Infections
CDC	Center for Disease Control and Prevention
BUMC	Banner University Medical Center
MLST	Multi-locus sequence typing
CLSI	Clinical and Laboratory Standard Institutes
PaLoc	Pathogenicity locus
RT	Ribotype

The Gram-positive and spore-forming anaerobic bacterium *Clostridioides difficile* (formerly named *Clostridium difficile*) is a leading cause of antibiotic-associated diarrhea that may be self-limiting, or progress to severe and fulminant (pseudomembranous) colitis or toxic megacolon[1–4]. There has been an increased incidence of *C. difficile* infection (CDI) over the past two decades[5–8] and, in the USA, this coincides with the emergence and spread of ribotype 027 strains [also called RT027 or BI or NAP1 based on the phylogenetic test[9,10]]. While RT027 remains the most prevalent healthcare-associated *C. difficile* ribotype, its frequency has been steadily declining[11]. Multiple surveillance studies indicate a changing trend in the *C. difficile* ribotype frequency distribution, particularly the emergence of RT106 (also called Group "DH" or "NAP11") in regions where it was previously rarely

[1]School of Animal and Comparative Biomedical Sciences, The University of Arizona, Tucson, AZ, USA. [2]Department of Pediatrics, The University of Arizona College of Medicine, Tucson, AZ, USA. [3]Department of Immunobiology, The University of Arizona, Tucson, AZ, USA. [4]Bio5 Institute for Collaborative Research, The University of Arizona, Tucson, AZ, USA. [5]Southern Arizona VA Health Care System, Tucson, AZ, USA. [6]School of Animal and Comparative Biomedical Sciences, University of Arizona, 1117 E Lowell St, Bldg. 90, Room 227, Tucson, AZ 85721, USA. [7]These authors contributed equally: Bryan Angelo P. Roxas and Jennifer Lising Roxas. ✉email: gayatri@arizona.edu

found. In 2008, RT106 was second to RT027 as the most dominant ribotype in England, and was also identified in neighboring European countries including Spain and Ireland[12–14]. However, during the same period, RT106 was rarely identified elsewhere in Europe, or in the USA and Canada[15], where RT027 and RT014/020 were predominant[13,15]. By 2012, RT106 emerged as the second most dominant *C. difficile* molecular type in the ten US states participating in the Centers for Disease Control and Prevention (CDC) Emerging Infections Program (EIP) surveillance[16]. From 2014 to 2017, RT106 replaced RT027 as the most prevalent ribotype recovered from community-associated CDIs[16–21].

Currently, Arizona is not a participant in the CDC EIP program, and no molecular typing data or epidemiological trends are available for this state. As part of an ongoing surveillance to rectify this gap in knowledge, we determined the ribotype frequency of *C. difficile* isolates recovered from patients at a tertiary University Medical Center in Tucson, Arizona between August 2015 and July 2018. Consistent with broader trends in the country, we noted increased prevalence of RT106 strains in our patient population. Since little is known about these strains[22], we focused on genomic and phenotypic characterization of all recovered RT106 isolates with the goal of identifying genetic factors contributing to the increased prevalence of this molecular type.

Results

Clostridioides difficile RT106 is the second-most prevalent molecular type in an acute-care teaching hospital in Tucson, Arizona. From August 2015 to July 2018, we recovered 788 *C. difficile* isolates from adult patients confirmed to be CDI-positive via a PCR test (employed until February 2017) or a "two-step" GDH/EIA test [Glutamate Dehydrogenase (assesses live *C. difficile*); Enzyme Immunoassay (detects *C. difficile* glycosyltransferase toxins TcdA and TcdB)] employed from March 2017. To ensure test-result consistency, we first verified the presence of *tcdB*, the same gene assayed in the PCR test, in all samples collected from March 2017 to July 2018. Overall, 519/788 isolates contained *tcdB* or expressed EIA-detectable levels of TcdA/B. Ribotype analysis revealed a diversity of strains in the patient population, with RT027 being the most frequently isolated strain (n = 144) (Fig. 1). RT106 (n = 38) was the second most frequently identified ribotype over the 3-year period.

RT106 isolates are virulent in an animal model of infection. Prior to detailed characterization of RT106 isolates, we verified the virulence of the representative strain GV599 in the Golden Syrian hamster model of acute *C. difficile* infection. All infected animals succumbed to disease within 6 days of spore inoculation (Supplemental Fig. S1a). Microscopy-based visualization of colonic tissue sections revealed classic *C. difficile* infection pathology including gross hemorrhage, epithelial erosion and inflammatory infiltrates (Supplemental Fig. S1b).

RT106 strains harbor one clade-specific novel genetic element. Whole genome sequencing was performed on all 38 RT106 strains recovered in our surveillance (Supplemental Table S2), and data were compared to 1425 publicly available *C. difficile* strain sequences. Based on single nucleotide polymorphism (SNP) analyses[23,24], the strains were not clonal, and the two closest-related isolates (GV597 and GV753) were divergent by 113 SNPs. Overall, RT106 genomes were most-closely related to RT002 strains[25].

Our 38 RT106 strains mapped closely to 33 previously sequenced RT106 strains from pediatric patients[26,27] and 23 other strains of unknown ribotype (highlighted in red in Fig. 2b). Evolutionary analysis of the 94 strains containing the entire GI1 was performed using MEGA X (Fig. 2c). We performed in silico ribotyping on the 23 strains, and 13/23 (those with currently available closed genome sequence) generated a clear RT106 PCR fragment pattern. For an additional assessment of genome relatedness, we performed in silico Multi-Locus Sequence Typing (MLST) on all 94 strains; this method differentiates organisms into Sequence Types [STs[28]]. 92/94 strains were sequence type ST42, whereas 2/94 belonged to the closely-related sequence type ST28[29]. Taken together, all 94 strains interrogated in these analyses grouped together in a distinct RT106 clade (Fig. 2b,c)[30].

Up to three unique genomic islands GI1, GI2 and GI3 are associated with the RT106 clade (Fig. 2a), and GI1, a novel 46 kb element reported for the first time herein, is invariantly carried by all RT106 strains. GI1 and GI3 were also predicted as genomic islands in our analysis of an RT106 strain BR81 genome using IslandViewer 4, which used SIGI-HMM and IslandPath-DIMOB as horizontal gene transfer predictors[31–34]. GI2 (also 46 kb) was previously identified in RT106 strains recovered from pediatric patients[27], and its overall prevalence in the RT106 clade is 7.4% (7/94 strains). GI3 (a 29.4 kb element) prevalence is 13.8% (13/94 strains). GI1 has features of conjugative mobile genetic elements and contain DNA integration and transposition genes (Locus IDs FE556_11090, FE556_11095, FE556_11065, FE556_11085, FE556_11205, FE556_11240, FE556_11260, FE556_11275 in Supplemental Table S3). GI3 also contains genes associated with conjugative transfer (Locus IDs FE556_02435, FE556_02450, FE556_02470 in Supplemental Table S4). Genes predicted to encode anti-restriction modification, antibiotic-resistance and cell adhesion functions are also present in GI1 and GI3 (Fig. 2d; Supplemental Tables S3 and S4). No plasmid-like genes were found. All three islands display higher percentage GC content (38%, 45% and 37% for GI1, GI2 and GI3, respectively) than the rest of the *C. difficile* genome (28–29%).

Currently, the 46 kb GI1 appears to be uniquely and specifically associated with RT106 (Fig. 2b,c), and all sequenced strains belonging to this clade (38 from this study and 56 others identified in publicly available databases) harbor a complete GI1 island. GI1 has 99.91% pairwise identity among strains (100% GI1 identity in 48 strains; 44 strains with 1–2 SNPs; 2 strains with > 3 SNPs). Fragments of GI1 were, however, detected in some non-RT106 strains. GI2, previously identified in pediatric RT106 isolates[27], is present in only 1/38 adult RT106 strains from our surveillance (Fig. 2c); we also identified this island in the non-RT106 strain Y358 (GCF_00451525.2). The 29.4 kb GI3 is present in 8/38 of our adult RT106 strains, as well as 5 other RT106

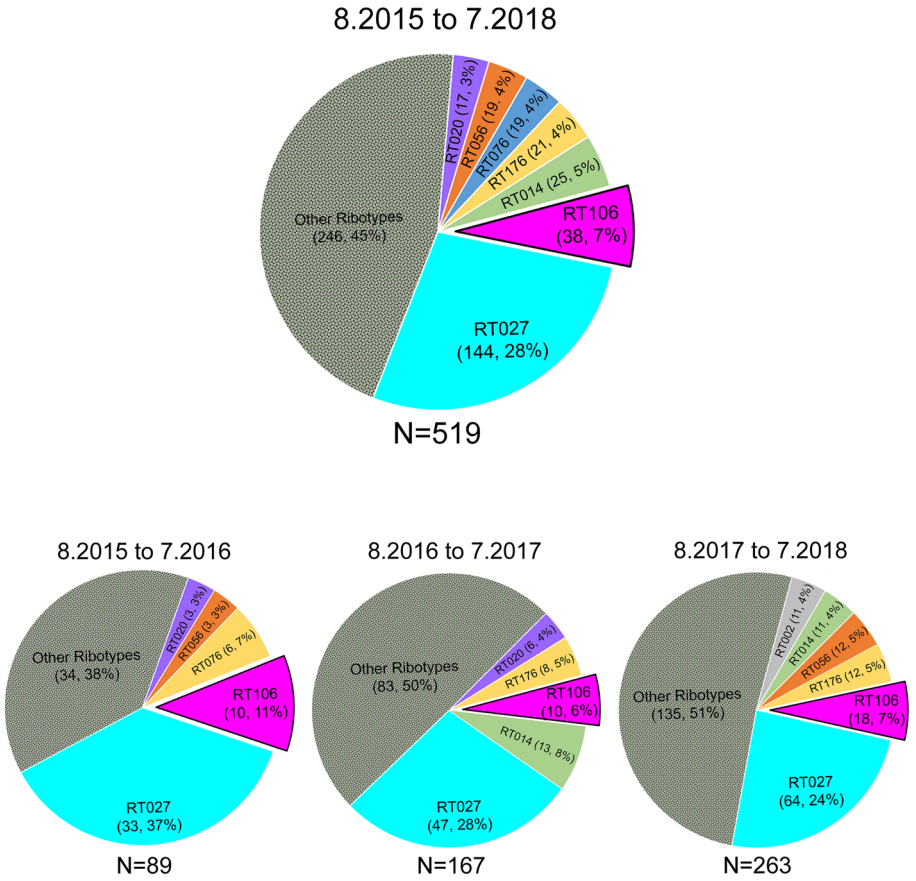

Figure 1. RT106 is the second most prevalent molecular type in a Tucson-area hospital. Top chart depicts ribotype distribution of 519 *tcdB* PCR-positive and/or TcdA/B ELISA-positive *C. difficile* strains from patient stool samples collected from August 2015 to July 2018 (8.2015 to 7.2018). Ribotype frequency and percent of total sample size are shown in parenthesis. Overall, RT106 is the second most frequently isolated molecular type, while RT027 is the most prevalent ribotype. Bottom charts depict ribotype distribution in 12-month periods. RT106 ranked second to RT027 as the most frequently isolated molecular type during 8.2015 to 7.2016 and 8.2017 to 7.2018. RT106 was the third most dominant ribotype during 8.2016 to 7.2017.

isolates in publicly available databases (Fig. 2c). We also identified GI3 in one non-RT106 strain (VRECD0053, GCF_900164815.1).

The 46 kb genomic island 1 is unique to RT106/ST42/ST28 strains. BLASTN analysis of the 46 kb GI1 against 1425 publicly available *C. difficile* genome sequences at the NCBI database resulted in the identification of 265 *C. difficile* strains that contain either segments (>7.7 kb, 98% identity) of or the entire genomic island. We concomitantly performed in silico MLST analysis to determine the respective sequence types, and then generated a maximum likelihood tree based on the core genome SNPs of 265 *C. difficile* strains harboring segments of or the entire G1 using Mega X[35]. GI1-related genes found in each strain were annotated based on gene function. Only RT106/ST42/ST28 strains harbor the complete 46 kb GI1, while other ST strains included in this analysis contain only shorter segments of the genomic island (Fig. 3).

A 7.1 kb gene segment (demarcated within a black dashed box; Fig. 3) is common to all MLST sequence type strains shown. SNP analysis was performed on the 7.1 kb gene segment to generate a molecular clock of GI1 via Mega-X[36] using maximum likelihood (ML) approach (Fig. 4). The molecular clock revealed gradual and progressive acquisition of gene elements in different strains, finally leading to an intact GI1. CD105KSE6, which branches most distantly from RT106 based on the alignment of the 7.1 kb segment of GI1, contained the least number of GI1-associated genes as opposed to STs branching closer to RT106/ST42/28.

To further interrogate whether GI1 was acquired via horizontal transfer, we compared the molecular clock of the 7.1 kb GI segment (Fig. 4) with the a molecular tree based on genes assumed to be refractory to horizontal gene transfer[37,38]. Thus, a minimum spanning tree was generated using the seven housekeeping genes utilized in MLST characterization to establish genetic relatedness of strains harboring the core 7.1 kb GI1 fragment (Supplemental Fig. S2). ST28, a sequence type that is included within the RT106 clade, is closely related to ST16, ST18 and ST46 based on sequence similarity of the core GI1 fragment (Fig. 4). However, only ST16, ST18 and ST28 are closely related based on the seven MLST gene loci; ST46 is distantly placed from ST16, ST18 and ST28 (Supplemental Fig. S2). A similar case is observed with the more predominant sequence type, ST48, within the

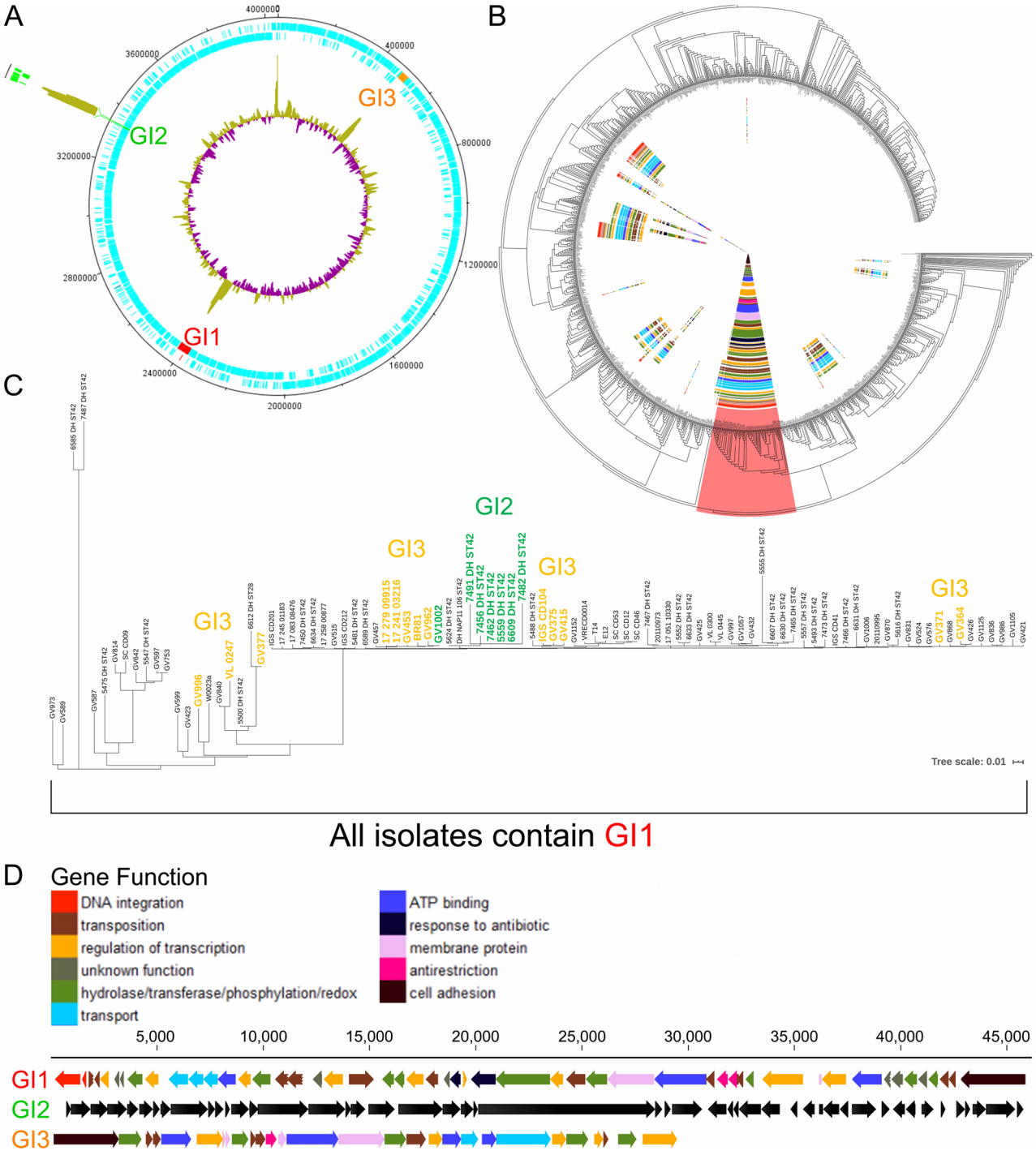

Figure 2. The RT106 clade may harbor up to three novel genomic islands. (**a**) Genetic islands GI1, GI2, and GI3 associated with RT106 are at three different locations in the genome. GV364, used as a representative genome, contains GI1 and GI3. Outer ring: Insertion site (green) of GI2 is shown relative to GI1 and GI3 locations; Blue indicate CDS (protein-coding DNA sequences). Inner ring: Purple denotes lower % GC compared to the overall % GC of the genome; Green denotes higher % GC compared to the overall % GC of the genome. Artemis DNA Plotter was used to generate genome circular map. (**b**) A composition vector tree of 1425 publicly available *C. difficile* and 38 RT106 genome sequences shows that RT106 strains clade together with 56 other strains (highlighted in red). (**c**) The relatedness of the 94 strains within the RT106 clade is shown in the maximum likelihood tree (log likelihood = – 29,380.67) based on the 3306 core SNPs identified using Panseq. Tree scale: 0.01 represents 0.01 substitutions per nucleotide site. Our clinical isolates are designated as "GV", whereas pediatric isolates from Chicago, Illinois[27] are designated "DH or ST". All 94 strains within the RT106 clade harbor the complete GI1. GI2 is present in 7 RT106 strains (green). Thirteen strains harboring GI3 (yellow) belong to 2 different subclades. (**d**) Gene arrangement, size and functions of GI1, GI2 and GI3. Functions of genes within GI1 and GI3 are named either via GO term or gene name. Genes within GI2 were previously identified and reported[27].

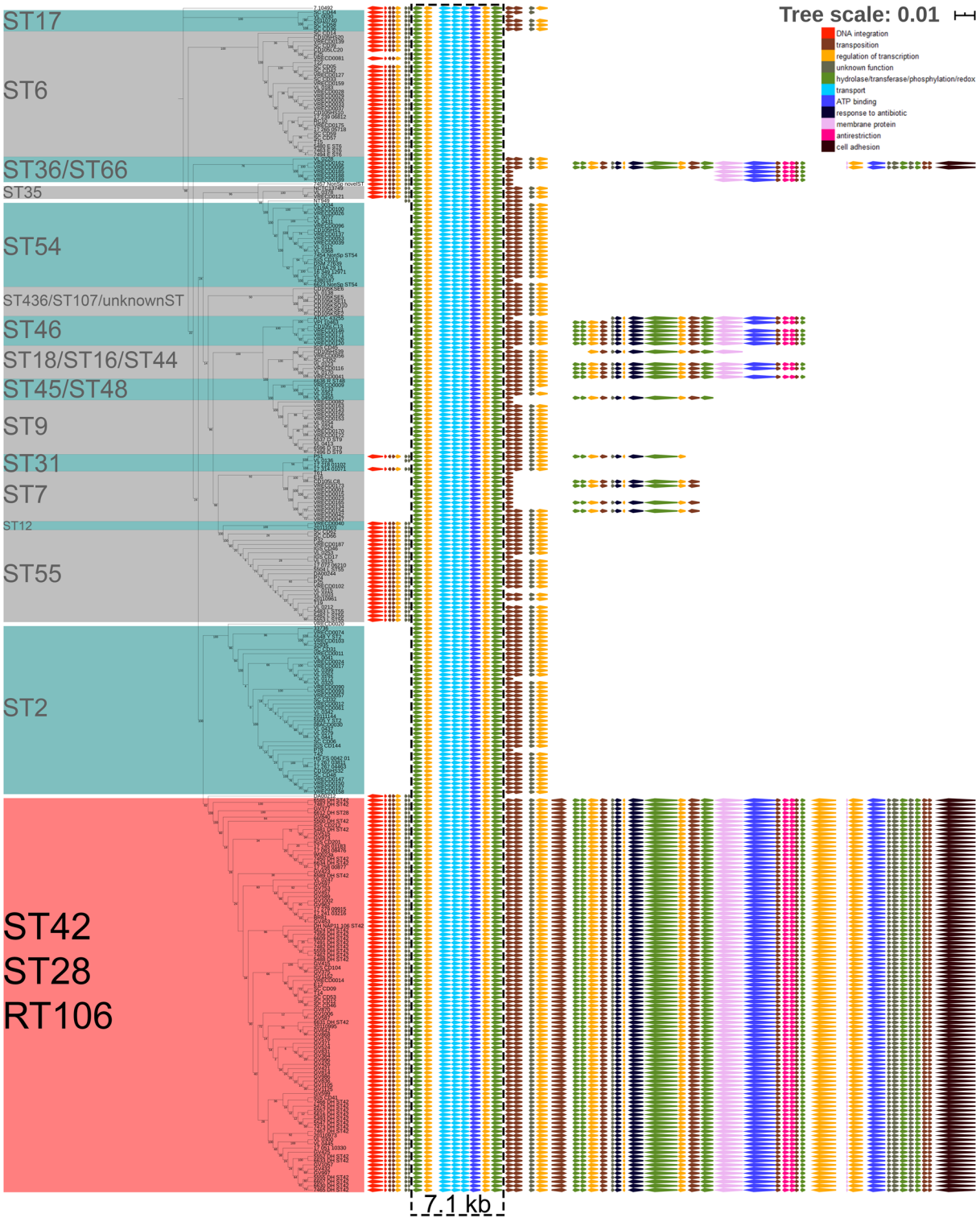

Figure 3. RT106 strains harbor a complete and unique 46 kb genomic island 1. The relatedness of the 265 *C. difficile* strains that carry GI1 segments (> 7.7 kb, 98% identity) is shown in a maximum likelihood tree (log likelihood = – 479,911.97) based on 40,879 core SNPs identified using Panseq. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (1000 replicates) are shown next to the branches. GI1 is drawn to scale on the right to illustrate regions present in different sequence types (ST). Tree scale: 0.01 represents 0.01 substitutions per nucleotide site. The complete 46 kb GI1 is present in RT106/ST28/ST42. Genes were colored based on functional categories from gene ontology (GO) analysis. A 7.1 kb region carried by all the strains (black dashed box) was used for determining progenitor STs of the element in the molecular clock analysis in Fig. 4.

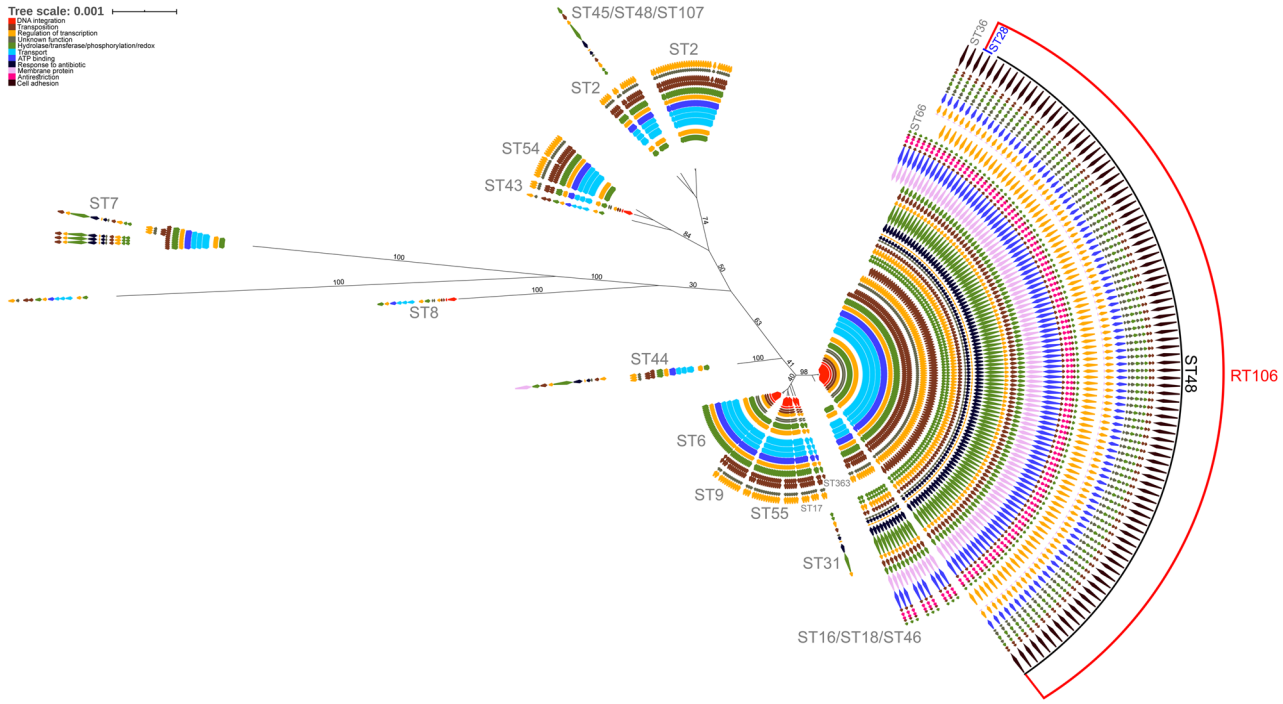

Figure 4. Molecular clock analysis reveals organization of the genomic island 1 via acquisition of distinct sub-elements. A timetree using the 7.1 KB consensus region in the 265 strains highlighted in black dashed box in Fig. 3 may offer clues towards the acquisition of sub-elements leading to the formation of GI1. Divergence times shown are relative times as no calibrations were provided. The estimated log likelihood value of the tree is − 14,227.79. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches. The phylogenetic tree is rooted using CD105KSE6. CD105KSE6 branches most distantly from RT106/ST28/ST42 clade based on alignment of the 7.1 kb region in GI1.

RT106 clade. ST48 is closer to ST42 and ST7 based on the seven housekeeping genes (Supplemental Fig. S2), and yet these ST strains map distantly in the core GI1-based molecular clock (Fig. 4). Since the tree topologies do not exhibit the same pattern, it is likely that GI1 is acquired laterally. Further analysis of tree topologies via likelihood ratio tests showed that the ML tree based on the 7.1 kb shared region within GI1 was significantly different from the ML tree based on core genome SNPs (P_{value} = 6.83E−74 calculated using approximately unbiased test) and the ML tree based on the seven MLST housekeeping genes (P_{value} = 1.17E−71).

The entire GI1 is not found in any other bacteria. However, two regions (8.4 kb and 13.7 kb) within GI1 were detected in other enteric bacteria (Fig. 5). The 13.7 kb gene segment was found in *Enterococcus faecium* EnGen0312 UAA407 at 99% sequence identity, while the 8.4 kb gene segment occurs in the same gene order but with some sequence plasticity in *Enterococcus faecium* EnGen0312 UAA407, *Anaerostipes hadrus* BPB5-Raf3-2-5, *Clostridium sporogenes* YH-Raf3-2-5 and *Roseburia intestinalis* M50/1 strains (89.8%, 90.4%, 90.5% and 92.2% DNA sequence identity, respectively).

Phenotypic characterization of RT106 isolates. Clade-specific properties, including those conferred by genes within GI1 could explain the emergence and spread of RT106 strains. Therefore, we assessed various virulence-associated phenotypes including antibiotic susceptibility, motility, toxin production, biofilm production and adhesion to collagen on the first 21 of the 38 RT106 strains chronologically obtained from our clinical surveillance.

RT106 strains display variable antibiotic susceptibility, with some isolates displaying multi-drug resistance. We determined the susceptibility of RT106 isolates to the antibiotics cefotaxime, vancomycin, erythromycin, clindamycin, levofloxacin, moxifloxacin, metronidazole, and tetracycline. All isolates were resistant to cefotaxime (minimum inhibitory concentration (MIC) > 32 mg/ml), but susceptible to vancomycin, metronidazole, and tetracycline (Table 1). 18/21 strains had intermediate resistance to clindamycin (MIC = 4–6 mg/ml). Three isolates (GV371, GV423, GV432) were highly resistant to erythromycin (MIC > 256 mcg/ml). Clindamycin and erythromycin belong to the macrolide-lincosamide-streptogramin B (MLS_B) group of protein synthesis inhibitors. MLS_B resistance in *C. difficile* has been associated with the acquisition of *erm* genes[39] or nucleotide substitution (C → T) at position 656 within the 23S rDNA[40]. None of the RT106 strains harbor the *erm* genes, while only GV415 had the 23S rDNA 656C>T substitution (Supplemental Table S1); however, GV415 has low-level resistance to clindamycin (MIC = 4 mg/ml) and is susceptible to erythromycin.

All RT106 isolates, except GV597, were susceptible to the fluoroquinolone levofloxacin. GV597, GV453, GV587, and GV642 had intermediate resistance to the fluoroquinolone moxifloxacin (MIC = 4–6 mg/ml).

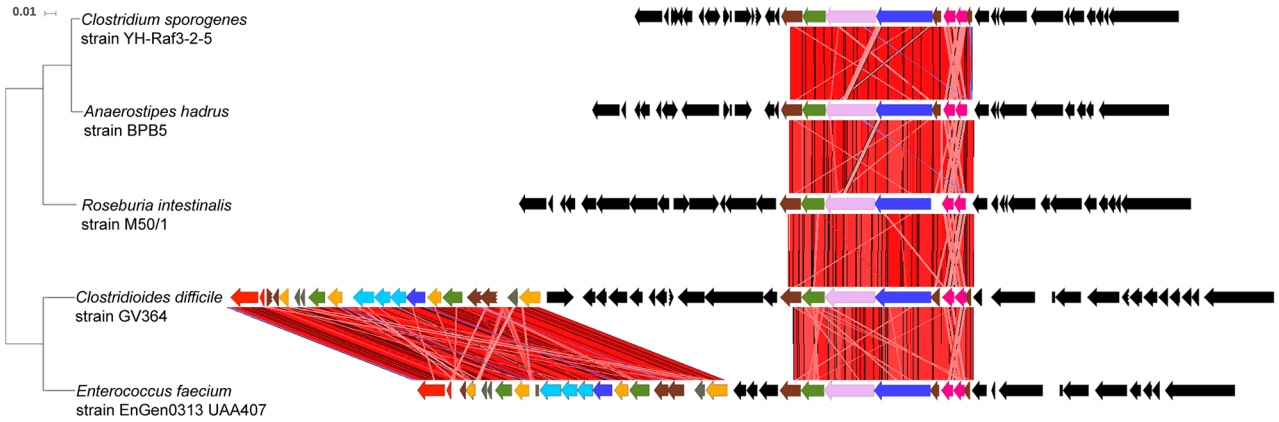

Figure 5. Human commensal microbiota may contribute to the acquisition of the 46 kb genomic island 1. The complete 46 kb GI1 is not present in any other microbial genome or plasmid sequence, but two gene segments (8.4 kb and 13.7 kb) within the island are found in other human enteric bacteria. The 8.4 kb gene segment is present in *Enterococcus faecium* EnGen0312 UAA407, *Anaerostipes hadrus* BPB5-Raf3-2-5, *Clostridioides sporogenes* YH-Raf3-2-5 and *Roseburia intestinalis* M50/1 strains (89.8%, 90.4%, 90.5% and 92.2% DNA sequence identity, respectively). *E. faecium* also harbors a 13.7 kb gene segment at 99% sequence identity. These two gene segments are found in *E. faecium* as part of a 36 kb genomic element. RT106 strains do not carry this 36 kb genomic element, but other *C. difficile* strains (VL0228 and 17-314-01071) strains have the identical 36 kb genomic element. Tree scale: 0.01 represents 0.01 substitutions per nucleotide site.

However, these fluoroquinolone-resistant RT106 isolates do not encode mutations in GyrA (T82I, T82V, D71V, D81N and A118T) or GyrB (D426V, D426N, R447L, R447K, S366A and S416A) associated with fluoroquinolone resistance[41–45] (Supplemental Table S1). The levofloxacin-resistant GV597 strain harbors an A421T mutation within the primary dimer interface of the conserved topoisomerase domain of gyrase A (Supplemental Table S1), but GyrA A421T mutation is not previously known to be associated with fluoroquinolone resistance in *C. difficile*.

GI1 harbors a gene encoding a VanZ family protein (locus ID FE556_11215; Supplemental Table S3). VanZ family proteins were previously implicated in teicoplanin resistance[46]. The GI1-encoded *vanZ* gene, present in all RT106 strains, is not found within *C. difficile* strains 630, VPI and BI-1 (Supplemental Table S1). Consistent with this, all RT106 isolates exhibit modest increase in resistance to teicoplanin compared to reference strains (T7, BI-1, 630, VPI; Table 1); the teicoplanin CLSI and EUCAST breakpoint values for *C. difficile* have not been established. Cultivation of RT106 strains in sub-inhibitory concentration (MIC) of teicoplanin (0.0125 mg/mL) resulted in increased teicoplanin resistance in 7/21 strains (Supplemental Table S5).

RT106 strains display collagen-dependent biofilm formation. Biofilm formation could facilitate intestinal colonization and persistence, and possibly contribute to recurrence[47]. RT106 strains display variable biofilm densities on an abiotic plastic surface (Fig. 6a). Since GI1 encodes a putative SrtB-anchored collagen-binding adhesin (locus ID FE556_11350; Supplemental Table S3), we tested the ability of RT106 strains to form biofilms on type I and type III collagen, the major collagen types present in the extracellular matrix of normal human intestines[48].

Biofilm densities of the non-RT106 toxigenic *C. difficile* strains BI1, 630 and VPI did not increase in the presence of collagen (Fig. 6b). However, eleven RT106 strains displayed collagen-dependent increase in biofilm formation when cultured on wells coated with both type I and type III human collagen. Overall, RT106 strains, as a group, have increased likelihood of displaying collagen-dependent biofilm formation.

We also interrogated the ability of the strains to form biofilms on either human type I or type III collagen individually. Although some RT106 strains showed increased biofilm formation on either collagen type (6 to human type I collagen; 5 to human type III collagen) (Supplemental Fig. S3), the RT106 strain group did not show collagen-dependent biofilm formation when only one collagen type was used for collagen coating. Curiously, GV426, GV453, and GV457 showed synergistic increase in biofilm formation to human types I and III collagen (Fig. 6).

We also tested the ability of the 21 RT106 strains to form biofilms on rat type 1 collagen and found that ten strains formed denser biofilms on rat collagen (Supplemental Fig. S4). GV425, GV426, GV432, GV453 and GV457 consistently formed denser biofilms on human and rat type I collagen compared to uncoated wells.

RT106 strains are variably motile. Flagella-dependent motility influences virulence of many pathogens[49]. All RT106 isolates tested, except GV375, GV415 and GV426, were motile (Fig. 7). We analyzed the genome of the non-motile RT106 strains for mutations in flagella-associated genes. In *C. difficile* 630 strain, flagella-associated genes are found in the F1 and F3 loci[50,51]. F1 and F3 loci are highly conserved in RT106; therefore, the nonmotile phenotype observed for GV375, GV415 and GV426 may possibly result from alterations in expression and/or post-translational modifications.

Strain	Minimum inhibitory concentration (MIC)								
	Teicoplanin	Cefotaxime	Clindamycin	Erythromycin	Levofloxacin	Moxifloxacin	Tetracycline	Vancomycin	Metronidazole
	0.016–256 mcg/mL	0.002–32 mcg/mL	0.016–256 mcg/mL	0.016–256 mcg/mL	0.002–32 mcg/mL	0.002–32 mcg/mL	0.016–256 mcg/mL	0.016–256 mcg/mL	0.016–256 mcg/mL
GV371	**0.125**	**>32**[a]	**4**[a]	**>256**[b]	4	3	0.50	0.75	0.50
GV423	**0.125**	**>32**[a]	**6**[a]	**>256**[b]	4	3	0.50	1.00	0.50
GV432	**0.125**	**>32**[a]	**4**[a]	**>256**[b]	4	2	0.09	1.00	0.38
GV597	**0.094**	**>32**[a]	**4**[a]	2	**12**[b]	**4**[a]	0.38	0.75	0.38
GV453	**0.125**	**>32**[a]	**4**[a]	1.50	4	**4**[a]	0.50	1.00	0.50
GV587	**0.125**	**>32**[a]	**6**[a]	1	4	**4**[a]	0.50	0.75	0.38
GV642	**0.125**	**>32**[a]	**4**[a]	1.50	4	**6**[a]	0.38	0.75	0.50
GV364	**0.125**	**>32**[a]	**4**[a]	1	6	2	0.50	1.00	0.50
GV375	**0.125**	**>32**[a]	**4**[a]	1	6	2	0.38	0.50	0.75
GV377	**0.125**	**>32**[a]	**4**[a]	2	4	3	0.38	0.75	0.75
GV415	**0.125**	**>32**[a]	**4**[a]	1	6	3	0.50	1.00	0.50
GV421	**0.125**	**>32**[a]	**4**[a]	1	4	2	0.38	0.75	0.75
GV425	**0.125**	**>32**[a]	**6**[a]	0.75	4	3	0.06	0.75	0.19
GV426	**0.094**	**>32**[a]	**4**[a]	1.50	6	3	0.50	0.75	0.50
GV524	**0.125**	**>32**[a]	**4**[a]	0.75	4	3	0.50	0.75	0.38
GV576	**0.125**	**>32**[a]	**4**[a]	1	4	3	0.38	0.75	0.50
GV589	**0.094**	**>32**[a]	**4**[a]	1	4	2	0.38	0.75	0.50
GV753	**0.125**	**>32**[a]	**6**[a]	1	4	3	0.38	0.75	0.50
GV599	**0.094**	**>32**[a]	3	1	4	3	0.38	0.50	0.38
GV457	**0.094**	**>32**[a]	3	1	4	3	0.13	0.75	0.38
GV515	**0.125**	**>32**[a]	3	1	4	2	0.38	1.00	0.50
630	0.064	**>32**[a]	**>256**[b]	**>256**[b]	4	3	**64**[b]	1.50	0.25
VPI	0.032	**>32**[a]	4[a]	1	3	1	0.38	1.00	0.25
T-7	0.064	**>32**[a]	3	1	4	3	0.50	1.00	0.50
BI-1	0.064	**>32**[a]	2	1	4	2	0.38	1.50	0.25
CLSI Breakpoints (mcg/mL)	NA	≥64 (Resistant)	≥8 (Resistant)	NA	NA	≥8 (Resistant)	≥16 (Resistant)	NA	≥32 (Resistant)
	NA	32 (Intermediate Resistance)	4 (Intermediate Resistance)	NA	NA	4 (Intermediate Resistance)	8 (Intermediate Resistance)	NA	16 (Intermediate Resistance)
	NA	≤16 (Susceptible)	≤2 (Susceptible)	NA	NA	≤2 (Susceptible)	≤4 (Susceptible)	NA	≤8 (Susceptible)
EUCAST Breakpoints (mcg/mL)	NA	NA	NA	NA	NA	NA	NA	≥2 (Resistant)	≥2 (Resistant)
	NA	NA		NA	NA	NA	NA	≤2(Susceptible)	≤2(Susceptible)

Table 1. Antibiotic susceptibility profiles of RT106 clinical isolates (this study). Numbers in bold font represent high MIC values. [a]Denotes that strain is moderately resistant to specific antibiotics. [b]Denotes that strain is highly resistant to specific antibiotics.

Most RT106 strains are robust toxin-producers. Toxigenic *C. difficile* produce up to two related glucosylating toxins, toxin A (TcdA) and toxin B (TcdB), which are encoded on the pathogenicity locus (PaLoc)[52,53]. Genome analysis of RT106 isolates revealed that all strains harbor the complete PaLoc and the gene for the TcdB1, instead of the highly toxigenic TcdB2 variant associated with select ribotypes including RT027[54,55]. We quantified secreted toxin, and observed that all RT106 strains, except GV457 and GV423, produced detectable TcdA/TcdB levels (Fig. 8). Nine RT106 isolates expressed TcdA/TcdB at levels comparable to the reference strain 630, while ten RT106 strains had similar (4/10) or higher (6/10) TcdA/TcdB levels compared to the RT027 strain BI-1.

Discussion

Consistent with broader trends in the United States, RT106 has emerged as the second leading ribotype from healthcare-associated cases in Southern Arizona[15–21]. Our genotypic and phenotypic characterization of multiple RT106 strains, along with the recent studies by Kociolek et al., represents an initial foray into defining key virulence properties of this clade[27,29].

The factors contributing to the emergence and expansion of RT106 strains are presently undefined, but they appear to be distinct from those postulated for the healthcare- and US-dominant RT027 clade. First, the enhanced ability of RT027 strains to utilize trehalose, a sugar increasingly used in food products since the early 2000s, may have provided a selective advantage for this clade, although this has recently been disputed[56,57]. None of the 94 sequenced RT106 genomes harbor the Leu-1721-Ile substitution in the TreR repressor or the four-gene insertion

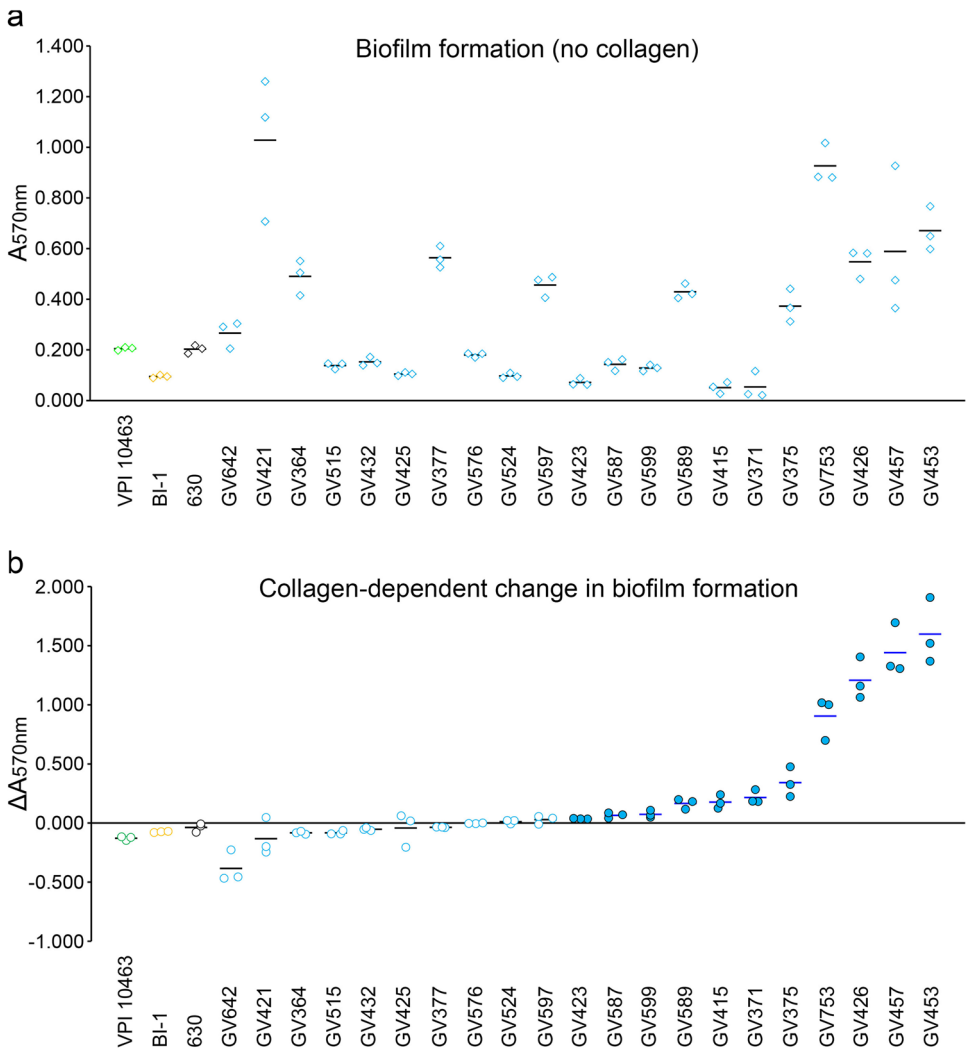

Figure 6. Clinical RT106 isolates display collagen-dependent biofilm formation. (**a**) 21 clinical RT106 strains (blue circles) and 3 non-RT106 toxigenic *C. difficile* strains (VPI, BI-1, and 630 designated as green, yellow and black circles, respectively) were cultured in uncoated or collagen-coated (combined types I and III) plastic wells for 72 h. RT106 strains displayed variable levels of biofilm on abiotic plastic wells. (**b**) Relative changes in biofilm densities (ΔA_{570nm}) were determined by comparing A_{570nm} of crystal violet-stained biofilms formed on human collagen (combined types I and III) vs. on uncoated plastic wells. Filled blue circles denote $P_{value} < 0.05$ determined using Student's t test to compare mean A_{570nm} by each strain on collagen-coated vs. uncoated wells. No difference in biofilm formation was observed when the reference *C. difficile* 630, BI-1 and VPI strains were cultured on wells with or without collagen. Overall, RT106 strains displayed denser biofilms on collagen-coated wells (One-sample one-tailed T-test; H_{alt}: mean $\Delta A_{570nm} > 0$; H_0: mean $\Delta A_{570nm} = 0$; $P_{value} = 0.02038$).

sequence that allow RT027 and RT078 strains, respectively, to grow on low levels of trehalose[56]. Still, our studies do not rule out unique sugar- or carbon source-utilization capabilities of RT106 strains.

Second, DNA gyrase mutations conferring fluoroquinolone resistance may have contributed to the emergence and spread of RT027 strains[42]. While RT106 isolates from the United Kingdom were highly resistant to moxifloxacin, those from North American surveillance studies, including ours, were mostly susceptible to fluoroquinolones (Table 1)[12,58–60]. Thus, fluoroquinolone resistance does not explain their emergence and spread in the United States.

Third, the PaLoc region of RT027 strains displays several key differences relative to the historic strain 630 (RT012)[61]. These include a point mutation in *tcdC* (though not in all isolates) that results in a truncated version of the anti-sigma factor TcdC, and expression of a variant of toxin B (TcdB2). TcdB2 has enhanced ability to enter host cells, is more cytotoxic, and exhibits wider tissue tropism[54,55]. In contrast to RT027 strains, the PaLoc of RT106 strains is 100% identical to 630[62–65]; thus, these strains encode full-length TcdC and express the TcdB1 toxin variant. Both RT027 and RT106 isolates produce variable amounts of TcdA/TcdB. Also, unlike 630 and RT106 strains, RT027 strains encode the binary toxin. Thus, toxin variations seem to be an unlikely driving force for the spread of RT106 strains.

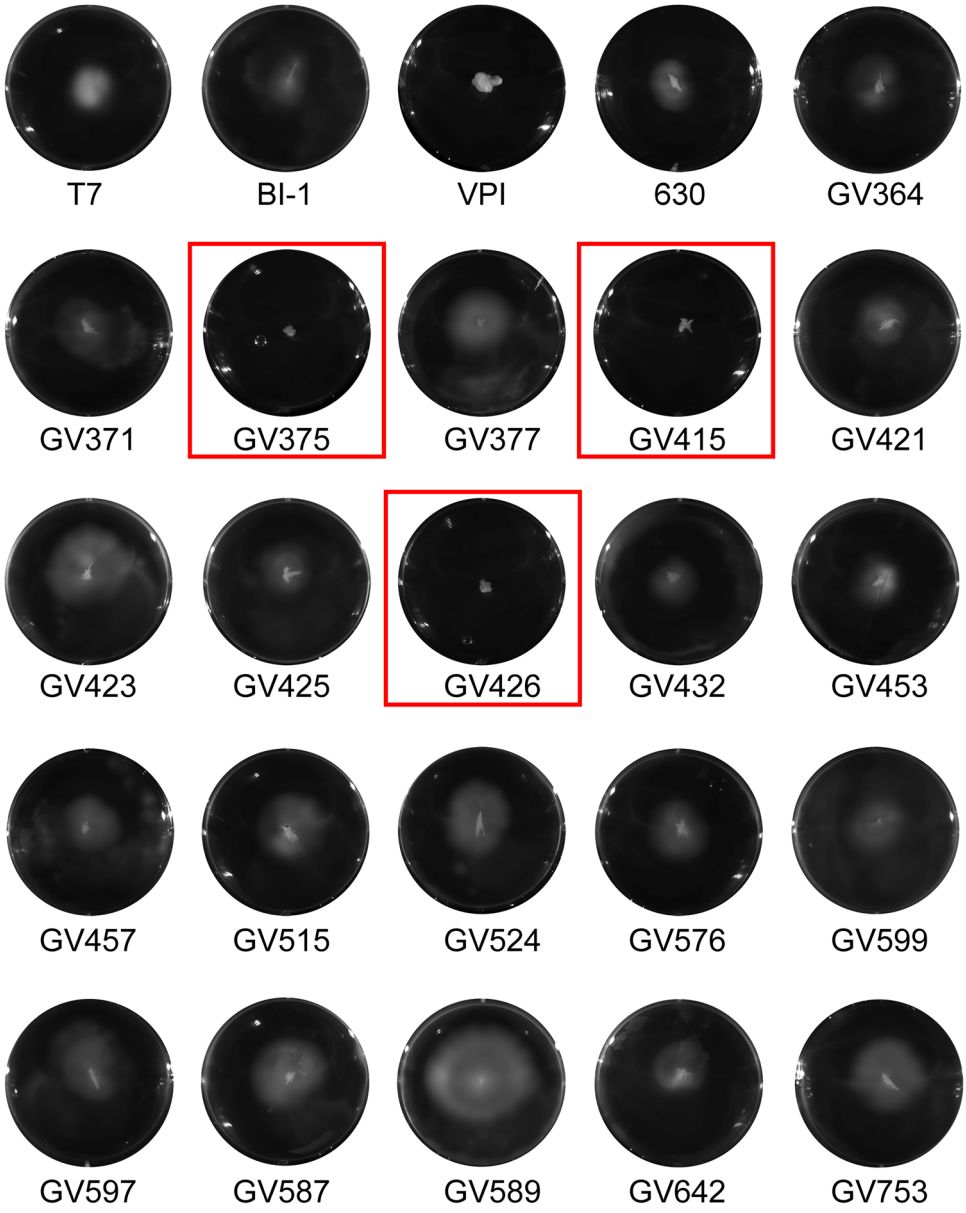

Figure 7. Clinical RT106 isolates are variably motile. All 21 clinical RT106 isolates, except GV375, GV415 and GV426 (red box), were motile in BHI soft agar. Motile (T7, BI-1, 630) and non-motile (VPI) reference strains are shown.

Detailed genome sequence analyses, however, suggest that the acquisition of novel genetic islands may be a contributor to RT106 emergence. All sequenced RT106 strains harbor a unique 46 kb genomic island (GI1) with a distinct GC content suggestive of horizontal acquisition. GI1 possesses several gene attributes that may confer competitive advantage to the RT106 clade. It harbors a *vanZ* allele (Locus ID FE556_11215), distinct from *vanZ1* (Locus ID FE556_05915; 49% identity) present elsewhere in RT106 genome and in other *C. difficile* strains including the well-studied 630[66] (Supplemental Table S1). In 630, VanZ1 was previously shown to confer low level resistance to the glycopeptide antibiotic, teicoplanin, but not to vancomycin[66]. The presence of a second VanZ allele may contribute to the modest increase in teicoplanin resistance of RT106 strains. The potential selective advantage of this phenotype cannot be ruled out; while teicoplanin is not FDA-approved for use in the US, it is widely used in Europe, Asia and South America.

In addition to the strain 630 cd2831 SrtB-anchored collagen-binding adhesin homolog (99% protein identity)[67], all RT106 strains encode a paralog within GI1 (locus ID FE556_11350; 33% protein identity with CD2831); this gene was earlier reported as an 'RT106-associated accessory gene'[29]. A subset of RT106 strains (13/94; 6 strains assayed for biofilm formation) also contains an additional paralog within GI3 (locus ID FE556_02390; 79% protein identity with GI1 locus ID FE556_11350). The robust collagen-dependent biofilm formation observed in the RT106 isolates may be due to the presence of any one or combination of these genes. Further

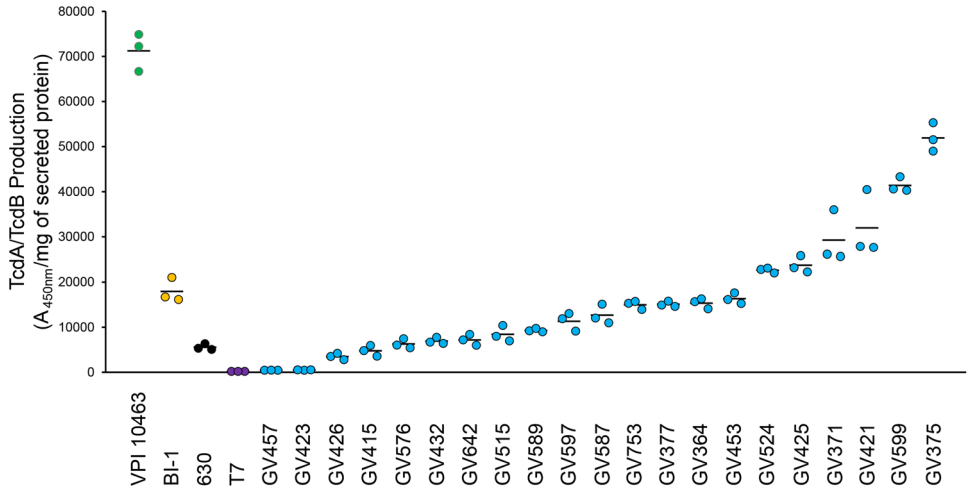

Figure 8. Most clinical RT106 isolates are robust toxin producers. All RT106 samples, except for GV457 and GV423, secrete similar or greater TcdA/TcdB levels compared to *C. difficile* 630 strain. Four strains (GV753, GV377, GV364, GV453) produced similar TcdA/TcdB levels as the BI-1 reference strain. Six strains (GV524, GV425, GV371, GV421, GV599, GV375) secrete more TcdA/TcdB compared to BI-1. TcdA/TcdB levels secreted after 72-h culture in BHI broth were normalized to mg of total secreted proteins. Mean A_{450nm}/mg of secreted protein and standard deviation are shown. Image is representative of two independent TcdA/TcdB ELISA assays with three sample replicates per condition.

investigation is required to parse the contribution of these genes to virulence. Since toxigenic *C. difficile* can breach the intestinal epithelium via cell junction disruption and/or epithelial cell death, thereby exposing the components of the extracellular matrix including collagen, strong vegetative cell and biofilm adhesion to collagen is a possible mechanism promoting *C. difficile* colonization of, and persistence in, the host.

GI1 also harbors genes for anti-restriction modification (*ardA*; Locus ID FE556_11265, FE556_11270), multi-drug resistance (*mfs*; Locus ID FE556_11225), methylglyoxal detoxification (*gloA*; Locus ID FE556_11330), and cation transport (Locus ID FE556_11135) containing a FieF domain (NCBI Conserved Domain cl30791) associated with iron-cobalt-zinc-cadmium resistance. Homologs of these genes are linked to virulence of other pathogens[68–70]. Finally, GI1 has features of a conjugative mobile element and contains genes for DNA excision/integration (Locus IDs FE556_11090 and FE556_11095) and encodes homologs of several proteins involved in Tcp conjugation machinery of *C. perfringens* including TcpE (YP_009063349.1; 46.24% similar to Locus ID FE556_11260), TcpG/TcpI hydrolase (YP_009063351.1; 51.49% similar to Locus ID FE556_11245), TcpF (YP_009063350.1; 49.35% similar to Locus ID FE556_11255), and TcpA (YP_009063346.1; 41.07% similar to Locus ID FE556_11305). It is presently unknown whether the entire 46 kb genomic island can mobilize to other *C. difficile* strains.

Fragments of GI1 were found in different *C. difficile* sequence types. Molecular clock analysis suggests that the complete island is a composite of sequences sequentially acquired from progenitor ST strains. The molecular clock based on the conserved GI1 segment is asynchronous with the one based on housekeeping genes (Fig. 4 and Supplemental Fig. S2). Further, consistent with higher GC content of GI1 relative to rest of the *C. difficile* genome, it is likely that the progenitor ST strains acquired the DNA segments from non-clostridial organisms via horizontal gene transfer. While the complete island is yet to be found in any other microbial genome or plasmid, two gene segments (8.4 kb and 13.7 kb) were detected in other enteric bacteria (Fig. 5). For the 8.4 kb gene segment, the most closely related sequences occur in *Roseburia intestinalis* M50/1 strains (92.2% identity). The 13.7 kb segment displays 99% identity to sequence within a 36 kb genomic island in *E. faecium*. While the 8.4 kb and 13.7 kb segment in GI1 may have been derived from *E. faecium*, the candidate donors of the other gene segments in GI1 are presently unknown. The presence of these genetic segments in disparate enteric organisms may suggest that they confer some selective advantage within the intestinal environment.

Conclusions

Clostridioides difficile RT106 is virulent in a hamster model of infection, and all sequenced isolates within this clade harbor a unique 46 kb GI1. Consistent with the presence of genes encoding a VanZ family protein and a SrtB-anchored collagen-binding adhesin within GI1, RT106 strains had increased teicoplanin resistance and robust collagen-dependent biofilm formation, respectively. Further investigation is required to implicate GI1 genes to RT106 virulence.

Methods

***Clostridioides difficile* surveillance.** This study, approved by the University of Arizona Institutional Review Board, utilized to-be-discarded stool specimens from diarrheic patients at the Banner University Medical Center (BUMC) in Tucson, Arizona between August 1, 2015 and July 31, 2018. Samples were collected and

stored at − 80 °C. From August 2015 to February 2017, *tcdB*-positive stool samples tested by BUMC via polymerase chain reaction (PCR) were included in the study. On March 2017, BUMC implemented the glutamate dehydrogenase (GDH) and toxin enzyme immunoassay for *C. difficile* testing. All GDH + samples were collected. We screened for the presence of *tcdB* in the GDH + /toxin− samples via PCR using the following primers: B1C (5′-GAAAATTTTATGAGTTTAGTTAATAGAAA-3′) and B2N (5′-CAGATAATGTAGGAAGTAAGT CTATAG-3′)[71]. For samples received during March 2017 to July 2018, only GDH+/toxin+ or GDH + /toxin− and *tcdB*-PCR-positive samples were analyzed in this study.

Ribotyping of clinical *C. difficile* isolates. Stool samples plated on taurocholate cycloserine cefoxitin fructose agar (TCCFA) were cultured anaerobically at 37 °C. Isolated colonies were lysed with G-Biosciences Toothpick-PCR, and supernatants were used as templates for ribotyping PCR using the following primers: 16S (5′-GTGCGGCTGGATCACCTCCT-3′) and 23S (5′-CCCTGCACCCTTAATAACTTGACC-3′)[72,73]. Isolated colonies were also submitted to the University of Arizona Genomics Core for genomic extraction using QIAGEN DNeasy column-based extraction kit and ribotyping PCR using the same 16S and 23S primers. PCR products were resolved via capillary electrophoresis using an AB Prism 3730 Genetic Analyzer (Applied Biosystems, Foster City, CA) and amplicon length evaluated using Marker 1.85 (SoftGenetics, State College, PA). Ribotype identification from electropherograms was determined using Webribo (https://webribo.ages.at/)[72].

DNA extraction. Genomic DNA samples were extracted using the protocol by Pospiech and Neumann[74], with modifications. Briefly, 50 mL overnight cultures of *C. difficile* were pelleted and resuspended in 5 mL of SET buffer (75 mM NaCl, 25 mM EDTA, 20 mM Tris, pH 7.5). Cell lysis was facilitated by adding lysozyme (5 mg/mL final concentration) and incubating samples at 37 °C for 30 min. 500 µL of 10% SDS and 25 µL of 100 mg/mL proteinase K were added, and samples were incubated at 55 °C for 2 h. 2.5 mL of 5 M NaCl and 5 mL of chloroform was added, and samples mixed with frequent inversions. Samples were centrifuged at 3000×*g* for 15 min, and aqueous phase was collected. DNA was precipitated using 1 volume of isopropanol. DNA was then spooled, transferred to a microfuge tube, rinsed with 70% ethanol, and vacuum dried.

Whole genome sequencing. DNA from 38 RT106 samples were submitted to the Office of Knowledge Enterprise Development (OKED) Genomics Core at Arizona State University (Tempe, Arizona, USA) for whole genome sequencing. Illumina-compatible genomic DNA libraries were generated on BRAVO NGS liquid handler (Agilent Technologies, Santa Clara, CA) using Kapa HyperPlus KK8514 library kit (Kapa Biosystems, Wilmington, MA). DNA was enzymatically sheared to approximately 600 bp fragments, end-repaired and A-tailed as described in the Kapa HyperPlus protocol. Illumina-compatible adapters with unique indexes (IDT #00989130v2; IDT technologies, Skokie, IL) were ligated individually on each sample. The adapter-ligated molecules were cleaned using Kapa pure beads (KK89002, Kapa Biosystems), and amplified with Kapa HiFi DNA Polymerase (KK2502, Kapa Biosystems). Fragment size of each library was analyzed using Agilent Tapestation, and quantified via qPCR using KAPA Library Quantification Kit (KK4835, Kapa Biosystems) and Applied Biosystems Quantstudio 5 Real-time PCR System before multiplex pooling and sequencing in a 2 × 250 flow cell on the MiSeq platform (Illumina, San Diego, CA) at the ASU OKED Genomics Core. Genomic libraries were split in 3 MiSeq runs. De novo genome assembly was performed using CLC Genomics Workbench 11 (QIAGEN Bioinformatics, Redwood City, CA). Depth of coverage ranges between 17X-608X (Supplemental Table S6). Contigs were annotated via Rapid Annotation using Subsystem Technology (RAST) Version 2.0[75–77]. Sequences for the 38 RT106 genomes were deposited through the National Center for Biotechnology Information (NCBI) Bankit (https://www.ncbi.nlm.nih.gov/WebSub/?tool=genbank) under the GenBank accession numbers listed on Supplemental Table S6.

Composition vector tree analysis. The 38 RT106 strains were mapped against a collection of all complete or draft *C. difficile* genomes sequences (1425 total sequences) available from the NCBI genome database (January 2019 download date). The composition vector tree was generated without sequence alignment by using a Composition Vector approach and CVtree Version 3.0[30]. Interactive Tree of Life v4.3 (https://itol.embl.de/)[78] was used to visualize and annotate the phylogenetic tree.

In silico multilocus sequence typing (MLST) and in silico ribotyping. Sequence types (ST) of *C. difficile* strains that claded with RT106 in the phylogenetic tree were determined based on the allelic patterns of 7 housekeeping genes[28] using the *C. difficile* MLST database (http://pubmlst.org/cdifficile). In silico ribotyping PCR analysis was performed on the uncharacterized strains using NCBI Primer-Blast[79] and the same 16S and 23S primers listed above. DH/NAP11/106/ST42 (Refseq assembly no. GCF_002234355.1), a complete closed genome, was used as a reference strain for the RT106 PCR fragment pattern.

Identification of RT106 genomic islands. A series of BLASTN searches (https://blast.ncbi.nlm.nih.gov/Blast.cgi)[80] was performed to identify the unique genetic elements associated with RT106. GV364 sequence was first compared to the complete closed genome sequence of *C. difficile* 630 strain (Refseq assembly no. GCF_000009205.2). Large genetic elements (> 10 kb) not found in *C. difficile* 630 were then compared to all 94 RT106 strain sequences to identify genetic elements associated only with RT106. The resulting genetic elements were verified to be unique to RT106 by performing BLASTN searches against 1425 publicly available *C. difficile* genome sequences at the NCBI database. Genome circular map of a representative strain GV 364 was generated

using Artemis DNA Plotter. GC content (%) and relative positions of GI1, GI2, and GI3 are indicated in the map (Fig. 2a).

Maximum likelihood (ML) trees of core genomes. ML trees were constructed for two groups of genomes; (1) 94 strains identified to clade together in the composition vector tree and found to contain a complete GI1, and (2) 265 strains that contain complete and partial (>7.7 kb and 98% identity) segments of GI1. Panseq[81] was used to determine the core SNPs. MEGA-X[36] was used to infer phylogenies by using the Maximum Likelihood method and Tamura-Nei model[82]. Trees with the highest log likelihood were shown in Figs. 2c and 3. The bootstrap consensus tree inferred from 1000 replicates was taken to represent the evolutionary history of the taxa analyzed[83].

Molecular clock analysis. The 7.1 kb genetic region common to 265 *C. difficile* strains was used to deduce the possible evolutionary formation of genetic island 1 on RT106. Mega-X[36] was used to construct a timetree inferred by applying the RelTime method[84,85] to the a phylogenetic tree whose branch lengths were calculated using the Maximum Likelihood (ML) method and the Tamura-Nei substitution model[82]. CD105KSE6 branched most distantly from RT106 (genetic distance of CD105KSE6 and GV973 = 0.024010404) based on the alignment of the 7.1 kb consensus region in GI1 and was used as the root for the tree. The bootstrap consensus tree inferred from 1000 replicates was taken to represent the evolutionary history of the taxa analyzed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test are shown next to the branches.

The independence of the acquisition of gene segments forming GI1 was tested by comparing the ML tree to a minimum spanning tree (MST) of MLST allele data profiles in the *C. difficile* MLST database (http://pubmlst.org/cdifficile). MST was created using PhyloViz v2.0[86].

Likelihood ratio test of tree topologies. Tree topologies were analyzed using IQ-TREE2[87] based on seven likelihood ratio-based tests (bootstrap proportion using RELL method test, one-sided Kishino-Hasegawa test, Shimodaira-Hasegawa test, weighted Kishino-Hasegawa test, weighted Shimodaira-Hasegawa test, expected likelihood weight, approximately unbiased test)[88–92] to compare ML trees based on core genome SNPs (designated as T_1 in this analysis) and seven MLST genes (designated as T_2) against the ML tree based on the 7.1 kb shared region within GI1 (designated as T_0) with the following hypotheses:

H_O: T_0 and T_1, or T_0 and T_2, would explain the sequence diversity of the 7.1 kb shared region within GI1 equally well ($T_0 = T_1$ or $T_0 = T_2$).

H_A: T_1 and/or T_2 does not explain the sequence diversity of the 7.1 kb shared region of GI1 ($T_0 \neq T_1$ or $T_0 \neq T_2$).

All tests were performed with 10,000 resamplings using the RELL method.

Antibiotic susceptibility testing. Overnight cultures of *C. difficile* strains were diluted to a McFarland scale of 0.5 (approximate $OD_{600nm} = 0.1$). 100 µL of the culture was plated onto Brucella blood agar. E-test strips (BioMerieux, Durham, NC) for the following antibiotics were applied on the agar: cefotaxime, vancomycin, erythromycin, clindamycin, levofloxacin, metronidazole, moxifloxacin, tetracyline and teicoplanin. Minimum inhibitory concentration, defined as the lowest concentration of the agent that inhibited bacterial growth, was determined. Antibiotic susceptibility was based on Clinical and Laboratory Standard Institute (CLSI) and European Committee on Antimicrobial Susceptibility Testing (EUCAST) breakpoints. There are no set standard breakpoints for teicoplanin. To test whether prior incubation with a sub-inhibitory concentration of teicoplanin promotes increased resistance, overnight cultures of *C. difficile* strains were diluted to a McFarland scale of 0.5 (approximate $OD_{600nm} = 0.1$). Five mL aliquots of the diluted culture were added into two new culture tube; Teicoplanin was added to one of the tubes to a final concentration of 0.0125 mg/mL. After 24 h of culture, antibiotic susceptibility testing was performed as indicated above.

Antibiotic resistance gene identification and profiling. Whole sequence genomes and proteomes were searched for antimicrobial resistance (AMR) genes using NCBI's AMRFinderPlus[93] and the Comprehensive Antibiotic Resistance Database's Resistance Gene Identifier Software Version 5.1.1 and Antibiotic Resistance Ontology Version 3.1.0[94]. Nonsynonymous SNPs in the AMR genes that may confer resistance to antibiotics used in susceptibility testing were compiled and tabulated in Supplemental Table S1.

Toxin ELISA. Relative levels of TcdA and TcdB toxins were determined using Alere Wampole A/B Toxin ELISA kit (Alere, Atlanta, GA). Overnight cultures of *C. difficile* strains were inoculated in 10 mL BHI at 1:100 dilution. Samples were cultured anaerobically for 72 h. Cultures were pelleted by centrifugation, and supernatants processed for Toxin ELISA following manufacturer's protocol and using BioTek Synergy automated plate reader. Total protein present in the supernatants were quantified using Pierce BCA protein assay kit. Relative amounts of toxin were normalized to total proteins.

Motility assay. Motility agar plates were prepared by adding 20 mL of BHI with 0.3% agar per well of a 6-well plate. *C. difficile* strains were cultured in BHI overnight. Approximately 5 µL of the culture was collected and stabbed into the motility agar. Plates were sealed and incubated in a humid, anaerobic chamber for 72 h, and then imaged using Bio-Rad ChemiDoc Touch Imaging System.

Biofilm assay. Twenty-four well plates were coated with: human or rat tail collagen type I (88 ng per well), human collagen type III (88 ng per well) or a combination of human collagen type I and type III (88 ng of each

collagen type per well). Overnight cultures of *C. difficile* strains were diluted in BHI containing 100 mM glucose ($OD_{600nm} = 0.1$). One mL of the culture was added per well of the uncoated or collagen-coated plate and incubated anaerobically for 72 h at 37 °C. Supernatants were removed gently by tilting plates onto a collection basin. Biofilms were washed twice by gently submerging plates in glass basins of PBS. Excess PBS was removed by inverting plates onto tissue paper. Biofilms were fixed for 20–40 min at 37 °C, and then stained with 1 mL of 0.2% filter-sterilized crystal violet for 30 min. Biofilms were washed twice with PBS as described above. For quantification of biofilm growth, 1 mL of 4:1 ethanol/acetone solution was added to each sample. 100 µL aliquots were transferred to a 96-well plate, and absorbance at 570 nm (A_{570nm}) was determined using BioTek Synergy automated plate reader. Relative changes in biofilm densities (ΔA_{570nm}) were determined by comparing A_{570nm} of crystal violet-stained biofilms formed on collagen-coated vs. on uncoated plastic wells.

***Clostridioides difficile* infection of Golden Syrian hamsters.** The Golden Syrian hamsters model[95] was employed to test GV599 virulence. A detailed protocol is included in the Supplementary Material. This animal study was approved by the Institutional Animal Care and Use Committee of the University of Arizona.

Ethical declarations. All methods were carried out in accordance with relevant guidelines and regulations. The *C. difficile* surveillance study was approved by the University of Arizona Institutional Review Board (Approval Number/ID 1707612129) as non-human subjects research. Informed consent was not required since to-be-discarded and de-identified stool samples were used.

Received: 9 April 2020; Accepted: 18 November 2020
Published online: 17 December 2020

References

1. Bartlett, J. G. Antibiotic-associated diarrhea. *N. Engl. J. Med.* **346**, 334–339 (2002).
2. McFarland, L. V. Antibiotic-associated diarrhea: epidemiology, trends and treatment. *Future Microbiol.* **3**, 563–578 (2008).
3. Nasiri, M. J. *et al. Clostridioides* (*Clostridium*) *difficile* infection in hospitalized patients with antibiotic-associated diarrhea: a systematic review and meta-analysis. *Anaerobe* **50**, 32–37 (2018).
4. Schroeder, M. S. *Clostridium difficile*-associated diarrhea. *Am. Family Physician* **71**, 921–928 (2005).
5. Ricciardi, R., Rothenberger, D. A., Madoff, R. D. & Baxter, N. N. Increasing prevalence and severity of *Clostridium difficile* colitis in hospitalized patients in the United States. *Arch. Surg.* **142**, 624–631 (2007).
6. Freeman, J. *et al.* The changing epidemiology of *Clostridium difficile* infections. *Clin. Microbiol. Rev.* **23**, 529–549 (2010).
7. Lessa, F. C. *et al.* Burden of *Clostridium difficile* infection in the United States. *N. Engl. J. Med.* **372**, 825–834 (2015).
8. Ho, J. *et al.* Disease burden of *Clostridium difficile* infections in adults, Hong Kong, China, 2006–2014. *Emerg. Infect. Dis.* **23**, 1671–1679 (2017).
9. McDonald, L. C. *et al.* An epidemic, toxin gene-variant strain of *Clostridium difficile*. *N. Engl. J. Med.* **353**, 2433–2441 (2005).
10. O'Connor, J. R., Johnson, S. & Gerding, D. N. *Clostridium difficile* infection caused by the epidemic BI/NAP1/027 strain. *Gastroenterology* **136**, 1913–1924 (2009).
11. Guh, A. Y. *et al.* Trends in incidence of long-term-care facility onset *Clostridium difficile* infections in 10 US geographic locations during 2011–2015. *Am. J. Infect. Control.* **46**, 840–842 (2018).
12. Brazier, J. S. *et al.* Distribution and antimicrobial susceptibility patterns of *Clostridium difficile* PCR ribotypes in English hospitals, 2007–08. *Euro Surveill.* **13**, 19000 (2008).
13. Bauer, M. P. *et al. Clostridium difficile* infection in Europe: a hospital-based survey. *Lancet* **377**, 63–73 (2011).
14. Wilcox, M. H. *et al.* Changing epidemiology of *Clostridium difficile* infection following the introduction of a national ribotyping-based surveillance scheme in England. *Clin. Infect. Dis.* **55**, 1056–1063 (2012).
15. Cheknis, A. *et al.* Molecular epidemiology of *Clostridioides* (*Clostridium*) *difficile* strains recovered from clinical trials in the US, Canada and Europe from 2006–2009 to 2012–2015. *Anaerobe* **53**, 38–42 (2018).
16. Center for Disease Control and Prevention. *2012 Annual Report for the Emerging Infections Program for Clostridium difficile Infection Infection* (Center for Disease Control and Prevention, Atlanta, 2012).
17. Centers for Disease Control and Prevention. *2013 Annual Report for the Emerging Infections Program for Clostridium difficile Infection Infection* (Centers for Disease Control and Prevention, Atlanta, 2013).
18. Centers for Disease Control and Prevention. *2014 Annual Report for the Emerging Infections Program for Clostridium difficile Infection Infection* (Centers for Disease Control and Prevention, Atlanta, 2014).
19. Centers for Disease Control and Prevention. *2015 Annual Report for the Emerging Infections Program for Clostridium difficile Infection Infection* (Centers for Disease Control and Prevention, Atlanta, 2015).
20. Center for Disease Control and Prevention. *2016 Annual Report for the Emerging Infections Program for Clostridium difficile Infection Infection* (Center for Disease Control and Prevention, Atlanta, 2016).
21. Centers for Disease Control and Prevention. *2017 Annual Report for the Emerging Infections Program for Clostridioides difficile Infection Infection* (Centers for Disease Control and Prevention, Atlanta, 2017).
22. Carlson, T. J., Blasingame, D., Gonzales-Luna, A. J., Alnezary, F. & Garey, K. W. *Clostridioides difficile* ribotype 106: a systematic review of the antimicrobial susceptibility, genetics, and clinical outcomes of this common worldwide strain. *Anaerobe* **62**, 102142 (2020).
23. Knetsch, C. W. *et al.* Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011. *Euro Surveill.* **19**, 20954 (2014).
24. Eyre, D. W. *et al.* Comparison of control of *Clostridium difficile* infection in six English hospitals using whole-genome sequencing. *Clin. Infect. Dis.* **65**, 433–441 (2017).
25. Wang, X. *et al.* Molecular typing of *Clostridium difficile*: concordance between PCR-ribotyping and multilocus sequence typing (MLST). *Open Forum. Infect. Dis.* **5**, S176–S176 (2018).
26. Kurka, H. *et al.* Sequence similarity of *Clostridium difficile* strains by analysis of conserved genes and genome content is reflected by their ribotype affiliation. *PLoS ONE* **9**, e86535 (2014).
27. Kociolek, L. K. *et al.* Whole-genome analysis reveals the evolution and transmission of an MDR DH/NAP11/106 *Clostridium difficile* clone in a paediatric hospital. *J. Antimicrob. Chemother.* **73**, 1222–1229 (2018).
28. Griffiths, D. *et al.* Multilocus sequence typing of *Clostridium difficile*. *J. Clin. Microbiol.* **48**, 770–778 (2010).

29. Kociolek, L. K., Gerding, D. N., Hecht, D. W. & Ozer, E. A. Comparative genomics analysis of *Clostridium difficile* epidemic strain DH/NAP11/106. *Microbes Infect.* **20**, 245–253 (2018).
30. Zuo, G. & Hao, B. CVTree3 web server for whole-genome-based and alignment-free prokaryotic phylogeny and taxonomy. *Genom. Proteom. Bioinf.* **13**, 321–331 (2015).
31. Bertelli, C. *et al.* IslandViewer 4: expanded prediction of genomic islands for larger-scale datasets. *Nucleic Acids Res.* **45**, W30–W35 (2017).
32. Waack, S. *et al.* Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. *BMC Bioinform.* **7**, 142 (2006).
33. Hsiao, W., Wan, I., Jones, S. J. & Brinkman, F. S. IslandPath: aiding detection of genomic islands in prokaryotes. *Bioinformatics* **19**, 418–420 (2003).
34. Hudson, C. M., Lau, B. Y. & Williams, K. P. Islander: a database of precisely mapped genomic islands in tRNA and tmRNA genes. *Nucleic Acids Res.* **43**, D48–D53 (2015).
35. Kumar, S., Stecher, G., Li, M., Knyaz, C. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
36. Knyaz, C., Stecher, G., Li, M., Kumar, S. & Tamura, K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.* **35**, 1547–1549 (2018).
37. Ravenhall, M., Škunca, N., Lassalle, F. & Dessimoz, C. Inferring horizontal gene transfer. *PLOS Comput. Biol.* **11**, e1004095 (2015).
38. Boc, A., Philippe, H. & Makarenkov, V. Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Syst. Biol.* **59**, 195–211 (2010).
39. Farrow, K. A., Lyras, D. & Rood, J. I. The macrolide-lincosamide-streptogramin B resistance determinant from *Clostridium difficile* 630 contains two *erm(B)* genes. *Antimicrob. Agents Chemother.* **44**, 411–413 (2000).
40. Schmidt, C., Loffler, B. & Ackermann, G. Antimicrobial phenotypes and molecular basis in clinical strains of *Clostridium difficile*. *Diagn. Microbiol. Infect. Dis.* **59**, 1–5 (2007).
41. Dridi, L., Tankovic, J., Burghoffer, B., Barbut, F. & Petit, J.-C. *gyrA* and *gyrB* mutations are implicated in cross-resistance to ciprofloxacin and moxifloxacin in *Clostridium difficile*. *Antimicrob. Agents Chemother.* **46**, 3418–3421 (2002).
42. Drudy, D., Kyne, L., Mahony, R. & Fanning, S. *gyrA* mutations in fluoroquinolone-resistant *Clostridium difficile* PCR-027. *Emerg. Infect. Dis.* **13**, 504–505 (2007).
43. Spigaglia, P. *et al.* Fluoroquinolone resistance in *Clostridium difficile* isolates from a prospective study of *C. difficile* infections in Europe. *J. Med. Microbiol.* **57**, 784–789 (2008).
44. Huang, H. *et al. Clostridium difficile* infections in a Shanghai hospital: antimicrobial resistance, toxin profiles and ribotypes. *Int. J. Antimicrob. Agents* **33**, 339–342 (2009).
45. Lin, Y.-C. *et al.* Antimicrobial susceptibilities and molecular epidemiology of clinical isolates of *Clostridium difficile* in Taiwan. *Antimicrob. Agents Chemother.* **55**, 1701–1705 (2011).
46. Arthur, M., Depardieu, F., Molinas, C., Reynolds, P. & Courvalin, P. The *vanZ* gene of Tn1546 from *Enterococcus faecium BM4147* confers resistance to teicoplanin. *Gene* **154**, 87–92 (1995).
47. Vuotto, C., Donelli, G., Buckley, A. & Chilton, C. *Clostridium difficile* biofilm. *Adv. Exp. Med. Biol.* **1**, 97–115 (2018).
48. Graham, M. F. *et al.* Collagen content and types in the intestinal strictures of Crohn's disease. *Gastroenterol.* **94**, 257–265 (1988).
49. Chaban, B., Hughes, H. V. & Beeby, M. The flagellum in bacterial pathogens: for motility and a whole lot more. *Semin Cell Dev Biol* **46**, 91–103 (2015).
50. Stabler, R. A. *et al.* Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains. *J. Bacteriol.* **188**, 7297–7305 (2006).
51. Stabler, R. A. *et al.* Comparative genome and phenotypic analysis of *Clostridium difficile* 027 strains provides insight into the evolution of a hypervirulent bacterium. *Genome Biol.* **10**, R102 (2009).
52. Hammond, G. A. & Johnson, J. L. The toxigenic element of *Clostridium difficile* strain VPI 10463. *Microb. Pathog.* **19**, 203–213 (1995).
53. Braun, V., Hundsberger, T., Leukel, P., Sauerborn, M. & von Eichel-Streiber, C. Definition of the single integration site of the pathogenicity locus in *Clostridium difficile*. *Gene* **181**, 29–38 (1996).
54. Lanis, J. M., Heinlen, L. D., James, J. A. & Ballard, J. D. *Clostridium difficile* 027/BI/NAP1 encodes a hypertoxic and antigenically variable form of TcdB. *PLoS Pathog.* **9**, e1003523 (2013).
55. Hunt, J. J., Larabee, J. L. & Ballard, J. D. Amino acid differences in the 1753-to-1851 region of TcdB influence variations in TcdB1 and TcdB2 cell entry. *mSphere* **2**, e00268-e1217 (2017).
56. Collins, J. *et al.* Dietary trehalose enhances virulence of epidemic *Clostridium difficile*. *Nature* **553**, 291–294 (2018).
57. Saund, K., Rao, K., Young, V. B. & Snitkin, E. S. Genetic determinants of trehalose utilization are not associated with severe *Clostridium difficile* infection outcome. *Open Forum. Infect. Dis.* **7**, 1 (2020).
58. Sundram, F. *et al. Clostridium difficile* ribotypes 027 and 106: clinical outcomes and risk factors. *J. Hosp. Infect.* **72**, 111–118 (2009).
59. Solomon, K. *et al.* PCR ribotype prevalence and molecular basis of macrolide–lincosamide–streptogramin B (MLSB) and fluoroquinolone resistance in Irish clinical *Clostridium difficile* isolates. *J. Antimicrob. Chemother.* **66**, 1976–1982 (2011).
60. Tenover, F. C., Tickler, I. A. & Persing, D. H. Antimicrobial-resistant strains of *Clostridium difficile* from North America. *Antimicrob. Agents Chemother.* **56**, 2929–2932 (2012).
61. Hunt, J. J. & Ballard, J. D. Variations in virulence and molecular biology among emerging strains of *Clostridium difficile*. *Microbiol. Mol. Biol.* **77**, 567–581 (2013).
62. Wüst, J., Sullivan, N. M., Hardegger, U. & Wilkins, T. D. Investigation of an outbreak of antibiotic-associated colitis by various typing methods. *J. Clin. Microbiol.* **16**, 1096 (1982).
63. Sebaihia, M. *et al.* The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.* **38**, 779–786 (2006).
64. Rupnik, M. & Janezic, S. An update on *Clostridium difficile* toxinotyping. *J. Clin. Microbiol.* **54**, 13–18 (2016).
65. Quemeneur, L. *et al. Clostridium difficile* toxoid vaccine candidate confers broad protection against a range of prevalent circulating strains in a nonclinical setting. *Infect Immun.* **86**, e00717–e00742 (2018).
66. Woods, E. C., Wetzel, D., Mukerjee, M. & McBride, S. M. Examination of the *Clostridioides (Clostridium) difficile* VanZ ortholog, CD1240. *Anaerobe* **53**, 108–115 (2018).
67. Hensbergen, P. J. *et al. Clostridium difficile* secreted Pro-Pro endopeptidase PPEP-1 (ZMP1/CD2830) modulates adhesion through cleavage of the collagen binding protein CD2831. *FEBS Lett.* **589**, 3952–3958 (2015).
68. Chakraborty, S., Gogoi, M. & Chakravortty, D. Lactoylglutathione lyase, a critical enzyme in methylglyoxal detoxification, contributes to survival of *Salmonella* in the nutrient rich environment. *Virulence* **6**, 50–65 (2015).
69. McMahon, S. A. *et al.* Extensive DNA mimicry by the ArdA anti-restriction protein and its role in the spread of antibiotic resistance. *Nucleic Acids Res.* **37**, 4887–4897 (2009).
70. Pitondo-Silva, A., Gonçalves, G. B. & Stehling, E. G. Heavy metal resistance and virulence profile in *Pseudomonas aeruginosa* isolated from Brazilian soils. *APMIS* **124**, 681–688 (2016).
71. Rupnik, M. *et al.* Characterization of polymorphisms in the toxin A and B genes of *Clostridium difficile*. *FEMS Microbiol. Lett.* **148**, 197–202 (1997).

72. Indra, A. *et al.* Characterization of *Clostridium difficile* isolates using capillary gel electrophoresis-based PCR ribotyping. *J. Med. Microbiol.* **57**, 1377–1382 (2008).
73. 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.* **175**, 261–266 (1999).
74. Pospiech, A. & Neumann, B. A versatile quick-prep of genomic DNA from gram-positive bacteria. *Trends Genet.* **11**, 217–218 (1995).
75. Aziz, R. K. *et al.* The RAST Server: rapid annotations using subsystems technology. *BMC Genom.* **9**, 75 (2008).
76. Overbeek, R. *et al.* The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). *Nucleic Acids Res.* **42**, D206-214 (2014).
77. Brettin, T. *et al.* RASTtk: a modular and extensible implementation of the RAST algorithm for building custom annotation pipelines and annotating batches of genomes. *Sci. Rep.* **5**, 8365 (2015).
78. Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Res.* **44**, W242-245 (2016).
79. Ye, J. *et al.* Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinform.* **13**, 134 (2012).
80. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
81. Laing, C. *et al.* Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions. *BMC Bioinform.* **11**, 461 (2010).
82. Tamura, K. & Nei, M. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. *Mol. Biol. Evol.* **10**, 512–526 (1993).
83. Felsenstein, J. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783–791 (1985).
84. Tamura, K. *et al.* Estimating divergence times in large molecular phylogenies. *Proc. Natl. Acad. Sci.* **109**, 19333–19338 (2012).
85. Tamura, K., Tao, Q. & Kumar, S. Theoretical foundation of the RelTime method for estimating divergence times from variable evolutionary rates. *Mol. Biol. Evol.* **35**, 1770–1782 (2018).
86. Francisco, A. P. *et al.* PHYLOViZ 2.0: providing scalable data integration and visualization for multiple phylogenetic inference methods. *Bioinformatics* **33**, 128–129 (2016).
87. Minh, B. Q. *et al.* IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol. Biol. Evol.* **37**, 1530–1534 (2020).
88. Kishino, H., Miyata, T. & Hasegawa, M. Maximum likelihood inference of protein phylogeny and the origin of chloroplasts. *J. Mol. Evol.* **31**, 151–160 (1990).
89. Kishino, H. & Hasegawa, M. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in Hominoidea. *J. Mol. Evol.* **29**, 170–179 (1989).
90. Shimodaira, H. An approximately unbiased test of phylogenetic tree selection. *Syst. Biol.* **51**, 492–508 (2002).
91. Shimodaira, H. & Hasegawa, M. Multiple comparisons of log-likelihoods with applications to phylogenetic inference. *Mol. Biol. Evol.* **16**, 1114–1114 (1999).
92. Strimmer, K. & Rambaut, A. Inferring confidence sets of possibly misspecified gene trees. *Proc. R. Soc. B.* **269**, 137–142 (2002).
93. Feldgarden, M. *et al.* Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates. *Antimicrob. Agents Chemother.* **63**, e00419–e00483 (2019).
94. Alcock, B. P. *et al.* CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. *Nucleic Acids Res.* **48**, D517-d525 (2020).
95. Vedantam, G. *et al.* An engineered synthetic biologic protects against *Clostridium difficile* infection. *Front. Microbiol.* **9**, 1 (2018).

Acknowledgements

This work was supported by funding from the National Institutes of Health [R33AI121590531(GV) and the US Dept. of Veterans Affairs [IK6BX003789(GV); I01BX001183(GV)]. Ribotyping work was performed with the support of the University of Arizona Genetics Core, University of Arizona, Tucson, AZ.

Author contributions

B.P.R. performed the comparative genomic analyses. J.L.R. performed phenotypic characterization and statistical analysis, analyzed data and drafted the manuscript. R.C.W., A.S.M. and F.A. isolated *C. difficile* from stool samples and performed ribotyping analysis. A.H. performed moxifloxacin and teicoplanin MIC determination and biofilm assay experiments. A.W., J.L. and S.J. conducted the pilot hamster experiment. SPE and KWS contributed in the conceptualization of the research and analyzed data. G.V. and V.K.V. conceptualized and funded the studies, finalized the manuscript and provided full project oversight. All authors read and approved the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-020-79123-2.

Correspondence and requests for materials should be addressed to G.V.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.