**Genomic study of the Type IVC secretion system in
Clostridium difficile: Understanding C. difficile evolution via
horizontal gene transfer**

Journal:	<i>Genome</i>
Manuscript ID	gen-2016-0053.R1
Manuscript Type:	Article
Date Submitted by the Author:	27-May-2016
Complete List of Authors:	<p>Zhang, Wen; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p> <p>Cheng, Ying; Chinese Center for Disease Control and Prevention</p> <p>Du, Pengcheng; Beijing Key Laboratory of Emerging Infectious Diseases</p> <p>Zhang, Yuanyuan; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p> <p>Jia, Hongbing; China-Japan Friendship Hospital,</p> <p>Li, Xianping; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p> <p>Wang, Jing; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p> <p>Han, Na; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p> <p>Qiang, Yujun; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p> <p>Chen, Chen; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p> <p>Lu, Jinxing; National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention/State Key Laboratory for Infectious Disease Prevention and Control</p>
Keyword:	Genome, Bacteria, Type IVC secretion system, Clostridium difficile, Genomic island

Draft

Genomic study of the Type IVC secretion system in *Clostridium difficile*: Understanding *C. difficile* evolution via horizontal gene transfer

**Wen Zhang^{1,2*}, Ying Cheng^{3*}, Pengcheng Du^{1,5,6*}, Yuanyuan Zhang^{1,5,6}, Hongbing Jia⁴,
Xianping Li^{1,2}, Jing Wang^{1,2}, Na Han^{1,2}, Yujun Qiang^{1,2}, Chen Chen^{1,5,6#}, Jinxing Lu^{1,2#}**

1 State Key Laboratory for Infectious Disease Prevention and Control, National Institute for Communicable Disease Control and Prevention, Chinese Center for Disease Control and Prevention, Beijing, China, 102206, **2** Collaborative Innovation Center for Diagnosis and Treatment of Infectious Diseases, Hangzhou, China, 310003, **3** Key Laboratory of Surveillance and Early-warning on Infectious Disease, Division of Infectious Disease, Chinese Center for Disease Control and Prevention, Beijing 102206, China, **4** Department of clinical laboratory, China-Japan Friendship Hospital, Beijing 100029, China, **5** Institute of Infectious Diseases, Beijing Ditan Hospital, Capital Medical University, **6** Beijing Key Laboratory of Emerging Infectious Diseases, Beijing 100011, China

* These authors contributed equally to this work.

Email: lujinxing@icdc.cn (JL); chenchen.bj2008@163.com (CC)

Running title: Type IVC secretion system in *C. difficile*

Abstract

Clostridium difficile, the etiological agent of *Clostridium difficile* infection (CDI), is a gram-positive, spore-forming bacillus that is responsible for ~20% of antibiotic-related cases of diarrhea and nearly all cases of pseudomembranous colitis. Previous data have shown that a substantial proportion (11%) of the *C. difficile* genome consists of mobile genetic elements, including 7 conjugative transposons. However, the mechanism underlying the formation of a mosaic genome in *C. difficile* is unknown. The type-IV secretion system (T4SS) is the only secretion system known to transfer DNA segments among bacteria. We searched genome databases to identify a candidate T4SS in *C. difficile* that could transfer DNA among different *C. difficile* strains. All T4SS gene clusters in *C. difficile* are located within genomic islands (GIs), which have variable lengths and structures and are all conjugative transposons. During the horizontal-transfer process of T4SS GIs within the *C. difficile* population, the excision sites were altered, resulting in different short-tandem repeat sequences among the T4SS GIs, as well as different chromosomal insertion sites and additional regions in the GIs.

Key words: Genome; Bacteria; Type IVC secretion system; T4SS; *Clostridium difficile*; Genomic island

Introduction

Clostridium difficile, the etiological pathogen of *Clostridium difficile* infection (CDI), is a gram-positive, spore-forming bacillus that is responsible for ~20% of antibiotic-related cases of diarrhea and nearly all cases of pseudomembranous colitis (Schwan 2009). Recent data have shown that *C. difficile* is the most common pathogen involved in healthcare-associated infections (HAIs), accounting for 12.1% of all HAIs in the United States (Huang et al. 2009). Due to its high morbidity and mortality, *C. difficile*-associated disease imposes a severe economic burden, estimated to cost the U.S. health care system in excess of one billion dollars annually (Drudy et al. 2006). In North America, Europe, and Asia, the prevalence of CDI has increased significantly and come into prominence in the last decade (DA 2013; Loo et al. 2006; Warny et al. 2005).

Previous findings have shown that *C. difficile* is a genetically diverse species, having a highly mobile and mosaic genome (He et al. 2010; Sebahia et al. 2006). Mobile genetic elements may contribute to the formation of the mosaic genome (Brouwer et al. 2011). For example, a relatively large proportion (11%) of the *C. difficile* 630 strain genome consists of mobile genetic elements, mainly in the form of conjugative transposons (CTns) (Sebahia et al. 2006). Several proven and putative CTns in *C. difficile* have been reported, such as CTns 1–7 in *C. difficile* 630 (Brouwer et al. 2011). Similar putative CTns also exist in 5 other sequenced *C. difficile* strains (BI1, BI9, 2007855, CF5, and M68) (Brouwer et al. 2012). Conjugative transposons are able to move from one bacterial cell to another through a process requiring cell-to-cell contact, which contributes to the spread of antibiotic-resistance and virulence genes in *C. difficile* (Brouwer et al. 2013).

The type IV secretion system (T4SS) is a versatile system that is essential for the virulence and even survival of some bacterial species (Brouwer et al. 2013; Dexi et al. 2012; Zhang et al. 2013). The

T4SS enables the secretion of protein and DNA substrates across the cell envelope. The T4SS was once believed to be the only secretion system to secrete DNA and to be present only in gram-negative bacteria. Previously, we identified a new subclass of T4SS, i.e., Type-IVC, which is present in the gram-positive genus *Streptococcus* (Zhang et al. 2012). In *S. suis* strain 05ZYH33, Type-IVC is located in a CTn (the 89K pathogenicity island), and can mediate the lateral transfer of this transposon to non-89K recipients (Li et al. 2011). In this study, we determined that this Type-IVC secretion system also exists in *C. difficile* and the horizontal transfer of T4SS GIs has occurred among *C. difficile* strains, based on genome-structure comparisons. We propose that the Type-IVC secretion system in *C. difficile* is responsible, at least in part, for the horizontal transfer of CTns and for the formation of its highly mobile, mosaic genome. Studying the function of the Type-IVC secretion system in *C. difficile* is useful for understanding how *C. difficile* acquires mobile genetic elements and clarifies the formation of its highly mosaic genome. This information would be useful for assessing the ability of *C. difficile* to acquire new antibiotic-resistance and virulence genes and for understanding their evolution.

Material and Methods

Bacterial strains used in this study

The *C. difficile* BJ08 strain was collected from a patient with diarrhea after long-term antimicrobial therapy in Beijing, 2008. Multilocus sequence typing (MLST) (Griffiths 2009), PCR ribotyping (O'Neill et al. 1996), and toxin detection (Kato et al. 1999) were conducted to investigate its molecular subtype and toxin profile. The BJ08 strain was defined as ST37, PCR ribotype (RT) 17, toxin A-negative, and toxin B-positive (A–B+).

A shotgun genome-sequencing method was used to obtain the genome sequence of the *C. difficile*

630 strain. Two DNA libraries containing 500-bp and 2-kbp DNA fragments were constructed for high-throughput sequencing on an Illumina Genome Analyzer IIx instrument, and 75-bp pair-end reads were collected. In total, 24,031,082 reads were generated with ~420-fold coverage of the *C. difficile* 630 genome (Sebaihia et al. 2006). The Illumina data were assembled using SOAPdenovo software (Li et al. 2010). The genome data were deposited into GenBank under Accession Number CP003939.

For genome comparisons, all available complete genome sequences of *C. difficile* (strains 630, 2007855, ATCC43255, CF5, M120, M68, R20291, CD196, and BI1) in the NCBI database were downloaded (Table S1). For the 6 unannotated strains (2007855, ATCC43255, BJ08, CF5, M120, and M68), the genes were predicted using Glimmer software (Delcher et al. 2007). Draft genome sequences of 16 *C. difficile* isolates were also downloaded from the NCBI database. Detailed information for these strains is shown in Table S1.

An additional 24 *C. difficile* isolates sampled from different countries were used in this study for PCR amplification and covered both toxin A+B+ and A-B+ strains (15/9), as well as 19 MLSTs (Table S2).

Genome comparisons and identification of T4SS GIs

To search for T4SS genes in the genomes of different *C. difficile* strains, we used T4SP software, which was described in detail in 2 of our previous papers (Zhang et al. 2012; Zhang et al. 2013). This program combines an alignment algorithm with protein-function predictions and domain evaluation, which helped to detect the candidate T4SS genes *virB1*–*virB11* and *virD4* (*VirB/D* genes). In this study, identification of a *VirB/D* cluster conformed to the following criteria: (1) the distance between 2 nearby *VirB/D* genes is less than 5 kb, (2) the total length of the *VirB/D* cluster is less than 50 kb, and (3) the

number of *VirB/D* genes in a *VirB/D* cluster is ≥ 3 .

To identify candidate genomic islands (GIs), the sequences of 10 *C. difficile* strains were compared using tblastx. All coding sequences in the query strain were located in the target genome sequence using the following parameters: E-value $\geq 1e-5$, identity ≥ 0.5 , and aligned length $\geq 50\%$. Only the best-matched hits were retained for multi-match results. Because of our criteria that all genes in an identified T4SS gene cluster as well as the neighboring regions were found in several *C. difficile* isolates, but not in other strains, it is possible that the T4SS clusters are located in GIs (referred to here as T4SS-type GIs). Based on the alignment results and analysis using the Sequencher program (Seiter 1992), the T4SS-type GIs and their precise locations within the genome were determined by synteny analysis between a genome with a *virB/D* cluster and one lacking a *virB/D* cluster. The function of genes in T4SS-type GIs was annotated using the NT, NR, Cluster of Orthologous Groups (COG), Kyoto Encyclopedia of Genes and Genomes (KEGG), and Swiss-Prot databases. To calculate the average nucleotide identity (ANI) value between the genomes of 10 *C. difficile* strains, we used ANItools (<http://ANI.bioinfo-icdc.org>) (Zhang et al. 2014).

PCR experiments

C. difficile strains were cultured on cycloserine-cefoxitin-fructose-egg yolk agar plates containing a cycloserine-cefoxitin supplement and 5% egg yolk and incubated anaerobically at 37°C for 48 h. *C. difficile* colonies were identified based on their characteristic morphologies and odor on agar plates, as well as their characteristics in gram-stain, latex-agglutination tests. All DNA from different isolates was ultimately identified by 16S rDNA and GDH gene amplification and sequencing. The primers used to amplify the CD630_04120 (VirD4) gene in T4SS GI1 of *C. difficile* strain 630 were designed using

DNASTar software, with the following sequences: 5'-TGCAAGATAAGGCAAAGTTTC-3' and 5'-ACTTCTGAAGCGTCTATCATATC-3'. PCR amplification cycle was performed by denaturation at 94°C (60 s), 55 °C annealing for 30 s, and extension at 72°C for 1 min. Each reaction was preceded by an initial denaturation step at 94°C for 5 min and terminated with a final extension step at 72°C for 5 min. To filter out false-positive results, all strains were tested with another pair of primers (5'-TCTTGCTAACGCAAACAGAAC-3' and 5'-AGTCCTCAAGGAGCTTGTAAT-3'). Only strains with positive results using both pairs of primers were defined as strains with the *VirD4* gene.

Phylogenetic analysis and GI sequence comparison

Multiple sequence alignments of the concatenated sequences of the *virB4*, *virB6*, and *virD4* genes were performed using MEGA4.0.2 software (Tamura et al. 2007). A phylogenetic tree was constructed using the neighbor-joining algorithm in MEGA4.0.2 software, and 1,000 subsets were generated for bootstrapping re-sampling analysis.

For genome comparisons with 10 strains, gene orthologs were determined using the OrthoMCL algorithm (Li et al. 2003). A matrix describing the genome contents was constructed with OrthoMCL, using a BLAST E-value cut-off of 1e-5 and an inflation parameter of 1.5. Genes included in all isolates were considered as core-genome genes. We examined SNPs through pairwise comparisons of 10 *C. difficile* isolates genomes, using the Mummer alignment program (Kurtz et al. 2004). Only SNPs located in core gene regions were retained. Phylogenetic trees based on 66,192 core-genome SNPs were constructed using the neighbor-joining algorithm in MEGA4.0.2 software (Tamura et al. 2007). Bootstrap was performed with 1,000 replicates. The methods used for detecting core genes and core SNPs were described previously (Chen et al. 2013). A phylogenetic tree based on the topoisomerase IA

gene was also built using MEGA4.0.2 software.

We used the Blastn method to compare pairs of 10 GI sequences with E-value cutoffs of 1.0, and only the results with alignment lengths ≥ 500 bp were retained for further analysis (Figure 2).

Results

Complete genome sequence of BJ08

We sequenced the genome of *C. difficile* BJ08 (ST37/RT17/A–B+) using the Illumina Genome Analyzer IIx system, following the manufacturer's instructions. The complete genome sequence is estimated to be 4,133,894 bp in size. We identified 3,461 Open Reading Frames (ORFs) larger than 300 bps that cover 81.6% of the genome, of which 77.5% matched to the COG database with an E-value of less than $1e-5$. Among these ORFs, we found that most genes are involved in common pathogenic pathways and comprise major virulence genes such as *tcdA* and *tcdB* at the pathogenicity locus (Du et al. 2014). The *tcdA* gene is truncated after 6310 bps, indicating that the BJ08 strain is A–B+. The genome sequence of BJ08 is the first *C. difficile* isolate from Asia to have its complete genome sequenced, which is helpful in studying CDI and understanding its evolutionary history and mode of spread worldwide.

Genome comparisons among *C. difficile* strains

Genome comparisons among 10 *C. difficile* genomes revealed that strain BJ08 has the highest ANI value (99.74%) when compared with *C. difficile* strain M68, whereas BJ08 has the lowest ANI value when compared with strain M120 (95.99%; Table S3). The ten *C. difficile* strains were found to encode 2475 core genes, and 66,192 SNPs were identified among these core genes. The phylogenetic

tree based on these core SNPs showed evolutionary relationships among the *C. difficile* strains, and the highest similarity was found to occur between strains BJ08 and M68 (Fig 1A).

The existence of T4SSs and their locations in mobile GIs

In this study, we found that 8 *C. difficile* strains have major components of the Type IVC secretion system. Three genes (*VirB4*, *VirB6*, and *VirD4*; previously identified in *S. suis*) clustered in *C. difficile* strains in the same orientation and have been shown to compose the core structure of the type-IVC secretion system (Zhang et al. 2012). This type-IVC secretion system serves to transport DNA among different strains. Among the 10 *C. difficile* strains studied, 10 T4SS gene clusters were identified in 8 strains. Only *C. difficile* CD196 and *C. difficile* BI1 did not carry a T4SS gene cluster. In contrast, *C. difficile* 630 has 3 T4SS gene clusters.

Further genome-comparison analysis revealed that the T4SS components were all located in GI regions. This type of GI is referred to here as a T4SS-type GI. The insertion sites and lengths of these 10 T4SS-type GIs were determined by performing comparisons with the BI1 genome sequence (Table 1 and Fig 1A). The length of the 10 T4SS-type GIs varied from 30–129 kb, which falls within the size range (10–200 kb) required for classification as a representative GI. All the T4SS-type GI sequences had significantly higher GC contents than the average GC percent in the respective genomes (Fig S1), which further suggests a foreign origin of these regions. At both ends of these T4SS GIs, we identified short direct repeat sequences (5–11 bp), which are also characteristic of CTNs. Among the 10 T4SS-type GIs studied, 8 were previously found to be CTNs in *C. difficile* (Brouwer et al. 2012; Brouwer et al. 2011). Two novel candidate CTNs were identified (T4SS GI5 in ATCC43255 and T4SS GI9 in BJ08) in this study (Table 1).

Although these T4SS-type GIs encode different genes and were of variable length, they shared same common characteristics (Fig 2). Based on genome comparisons and gene-annotation analysis, these T4SS-type GIs share several elements important for DNA transfer between strains, such as integrase, helicase, excisionase, and mobilization proteins. Similar to the 89K-GIs and other CTns, short direct-repeat regions were found at both ends of the T4SS-type GI regions. Six types of short repeat sequences in 10 T4SS GIs were identified, which ranged from 5 bp to 11 bp in length. Strains with the same short repeat sequences always inserted in the same location of the *C. difficile* genome (Fig 1A). For example, 3 GIs in A–B+ strains CF5, BJ08, and M68 have the same short direct-repeat region and insert in the same location (nucleotide position 466,669). GI 3 and GI 10 have the same repeat sequence that is inserted in the same location of the genome, as do GI4 and GI5. The existence of these short repeat regions provides a mechanism for self-circulation of the GI regions after splicing from the genome. With the help of T4SS, it was found that dsDNA, ssDNA, and GIs within the T4SS itself could transfer across the cell envelope during bacterial cell conjugation, as is known to occur with *S. suis*. The existence of T4SS-type GIs may explain why *C. difficile* has a highly mosaic genome with many GIs. Based on annotation results, these T4SS GIs also contain genes with other functions, such as DNA methylase, cell wall-associated hydrolases, topoisomerase IA, cell-surface proteins, ABC-type transport-system genes, transcriptional regulator, and a 2-component signal-transduction system (Fig 2, Table S4, and Table S5). Among these genes, we found that some functional genes are related to antibiotic resistance. For example, T4SS GI8 harbors 3 drug-resistance genes, M120GL000423 (*tet*), M120GL000409 (*aadE*), and M120GL000424 (*aadE*), where the first gene mediates tetracycline resistance and the other 2 genes mediate streptomycin resistance. In T4SS GI1 and GI10, the CD630_04340 and CDR20291_1779 genes, which are both annotated as genes encoding a Na⁺-driven

multidrug efflux pump, are also related with the drug-resistance mechanism.

Sequence comparisons also revealed similarities between GIs. For example, GI2 in *C. difficile* 630 is the shortest GI found in *C. difficile*, being only ~30 kb in size (Fig 2). GI2 has major T4SS components including *VirB4* (CD630_11100), *VirB6* (CD630_11120), and *VirD4* (CD630_11150), as well as other mobile elements (Fig 2). Although *VirB/D* genes share low similarity between species, *VirB4*, *VirB6*, and *VirD4* still showed 53%, 44%, and 57% similarity when compared with their counterparts in *S. suis*. These 3 genes are components of the core structure of the type IVC secretion system and potentially mediate the transport of their own and other DNA strands between strains. A 6-bp short repeat region “AATTTA” is located at both ends of the GI2 region, while the 89K T4SS GI of *S. suis* has a 15-bp repeat region. Both 89K and GI2 harbor integrase and excisionase (CD630_10910 and CD630_10920). Integrase is a site-specific recombinase that is presumably responsible for self-excision and integration of the GI into the bacterial chromosome (Li et al. 2011). The excision function of integrase is often simulated by the excisionase, both of which facilitate excision and inhibit integration (Sam et al. 2004). The protein encoded by CD630_11020 has a 180-aa C-terminal region homologous to *mobA* in 89K of *S. suis*.

During the transfer process of T4SS GIs within the *C. difficile* population, the excision sites could become altered, resulting in different short repeat sequences among T4SS GIs and different insertion sites in the chromosome. Our phylogenetic trees based on T4SS genes (Fig 2) and topoisomerase IA (Fig S2) genes both revealed that T4SS GI3, GI6, GI7, and GI9 are located in the same branch and have higher similarity than that observed with other GIs. However, GI3 has different short repeat sequences than those of GI6, GI7, and GI9. Detailed sequence analysis showed that GI3 also has the short repeat sequence “TGAGACGGTAG” found in GI6 at the 5' end. The change of excision sites

between GI3 and GI6 not only results in a new insertion site (Fig 1A), but also was associated with a 2-bp deletion at the 5' end and an additional 5.8-kb region at the 3' end (Fig 3).

Phylogenetic analysis of T4SS genes

The phylogenetic tree of genome-core SNPs (Fig 1A) revealed that the Type-IVC secretion system in *C. difficile* is mobile among strains with or without Type-IVC secretion systems, co-existing in several branches of the phylogenetic tree. Analysis of the other phylogenetic tree (Fig 2) obtained from the concatenated sequences of the *VirB4*, *VirB6*, and *VirD4* genes in *C. difficile* strains showed that the occurrence of these 10 T4SS-GIs was not caused by mutations, but by multiple DNA acquisitions from other strains. Three T4SS-GIs of *C. difficile* 630 (T4SS GI1, T4SS GI2, and T4SS GI3) were located in different branches. Thus, it is unlikely that they originated from the same ancestor and self-duplicated within a strain. Instead, they were likely inserted by different foreign DNA sequences in 3 independent events. T4SS genes of GI1, GI4, and GI10 were located in the same branch of the phylogenetic tree, while GI3, GI6, GI7, and GI9 were located in another branch. Strains in the same branch potentially share a common ancestor. The phylogenetic relationship among the 10 GIs studied is also supported by a phylogenetic tree based on the topoisomerase IA gene (Fig S2), which matched 100% with the T4SS gene tree.

The T4SS GIs located in the same branch typically showed high similarity in 1 region, but high divergence in other regions. For example, GI6 has an additional 5.8-kb region compared to GI3 at the 3' end, and GI9 has 2 large insertions near the 5' end of GI7. ATCC43255GL003458–ATCC43255GL003463 in GI5 of ATCC43255 were replaced by 4 genes (CD630_18650–CD630_18680) in GI3 of *C. difficile* 630. Five genes (CD630_18650–CD630_18690) in the same

location of GI3 were replaced by 2 genes (CF5GL000410 and CF5GL000411) in T4SS GI6 of the CF5 strain, although their neighboring left and right regions both showed high similarity (98% and 94%). The high-similarity regions among the T4SS GIs covered the T4SS genes in each case (Fig 2).

By comparing *C. difficile* T4SS gene sequences with entries in the NT database and identifying the best matches, we traced the candidate source of genes within the 10 T4SS GIs. The 10 T4SS GIs showed a clear mosaic structure, as the genes within the T4SS GIs had variable gene sources. As shown in Fig S3, most T4SS genes shared highest similarity with genes from *Streptococcus spp.* (except for the T4SS genes in GI2 and GI8), although the remaining genes had multiple originating sources. For example, a homologous sequence at the 3' end region in GI6 also exists in the ATCC *Sebaldella termitidis* 33386 strain. GI8, the largest T4SS GI being 129 kb in length, has an additional sequence inserted between M120GL000367 and M120GL000402 at the 5' end, which was also found in *Thermoanaerobacter* sp. X513.

Existence of T4SS in the *C. difficile* population

Using bioinformatics methods, we found T4SS genes in 6 of 16 (37.5%) *C. difficile* strains with draft genomes. Considering the possibility that T4SS genes are potentially located in the sequence gaps, the actual percentage of *C. difficile* strains with T4SS is potentially higher.

We also performed PCR experiments to determine the distribution of the *VirD4* gene in the *C. difficile* population, which is the most conserved gene among the 3 T4SS Vir genes. Similar to the genome-analysis results, the PCR results indicated that T4SS existed in several, but not all *C. difficile* strains. Among 24 *C. difficile* strains tested (Supplemental Table 2), 6 of 15 toxin A+B+ strains (40%) and 5 of 9 toxin A-B+ strains (55.6%) have *VirD4*-gene homologs (Fig 1B). In this study, 19 sequence

types were identified, among which 6 strains (ST1, ST37, ST48, ST54, ST55, and ST118) had VirD4-gene homologs, while the remaining 13 sequence type strains did not (Fig 1B).

Discussion

In this study, we identified 3 T4SS GIs (GI1, GI2 and, GI3) in *C. difficile* 630 that were previously demonstrated to be CTns (Ctn2, Ctn4, and Ctn5) (Brouwer et al. 2012; Brouwer et al. 2011). Using PCR and ClosTron retargeting technology, these GIs in *C. difficile* 630 were shown to become excised from the genome, form an extrachromosomal circular product, and then transfer to the recipient strain CD37 (Brouwer et al. 2012; Brouwer et al. 2011). The overall process is similar with that observed with the 89K GI in *S. suis*. Combined bioinformatics and functional analysis revealed that other T4SS GIs may also transfer between *C. difficile* strains in the same manner, since they share the same mobile genetic elements, such as integrase, the Type-IVC secretion system, and direct-repeat regions.

Our phylogenetic-analysis results suggested that the T4SS GIs may be mobile between strains (Fig 1A and Fig 2). Based on the phylogenetic tree generated from 66,192 core SNPs, strains with or without T4SS GIs can be located within the same branch (Fig 1A), which suggests that the T4SS GIs were not generated by a spontaneous mutation in an ancestor of this branch, but were potentially caused by horizontal gene transfer. For example, 4 strains (R20291, 2007855, BI1, and CD196) were located in the same branch and share a common ancestor. T4SS GIs were found in 2 strains (R20291 and 2007855), but not in the other 2 strains of the same branch (BI1 and CD196). The T4SS GI (GI7) in M08 shares the highest identity (100%) with GI6 in the CF5 strain, but not with GI9 in BJ08, which with highest similarity with M08 at the genome level. This finding also supports the hypothesis that T4SS GIs of *C. difficile* originate from horizontal genetic transfer between strains, rather than by

evolution of the whole genome.

Genome comparisons and gene-annotation work in this study both indicated that T4SS-type GIs in *C. difficile* and 89K-GIs in *S. suis* possess several similar elements important for DNA transfer between strains, such as integrase, helicase, excisionase, and mobilization proteins, as well as 2 short direct-repeat regions found at both ends of the GI regions. Thus, the horizontal transfer of T4SS GIs in *C. difficile* could occur in 5 steps, as was observed with GI 89K in *S. suis* (Chen et al. 2007; Li et al. 2011; Zhang et al. 2012). According to this model, in the first step, T4SS GIs are self-cleaved from the chromosome in the direct-repeat regions of T4SS GIs with the help of integrase and excisionase. The cleaved strand then forms an intermediate circle (Step 2) and transfers through the transport channel across the cell membrane via the activity of VirB6 (Step 3). VirB4 and VirD4 are ATPases that could provide the energy necessary for such transport. Cell wall-associated hydrolase functions to partially degrade the plasma membrane of bacteria, thereby reducing resistance to substrate secretion. The self-circularized T4SS GI sequences could be inserted into the chromosome of recipient cells (Step 4) and cause the formation of the mosaic *C. difficile* genome (Step 5). However, more experimental work needs to be performed to support this model.

T4SS GIs with the same short repeat sequences always inserted into the same chromosomal location, suggesting that the insertion sites of T4SS GIs in recipient cells are determined by their short repeat sequences. For example, T4SS GI3 and GI10 have “GTTGA” repeats at both ends and were inserted at the same location (Fig 1A), even though their structures are clearly different and their T4SS genes are located in 2 different branches of the evolutionary tree (Fig 2). T4SS GI4 and GI5 were inserted in the same location of the chromosome, as were T4SS GI6, GI7, and GI9 (Fig 1), and both groups carry the same short repeat sequences. Some strains without T4SS GIs still have 1 (instead of 2)

short repeat sequences at the same location in the genome. Thus, strains without T4SS GIs, such as CD196 and BI1, have the potential of incorporating T4SS GIs, as they also have the target sequence of the same short repeat sequences of T4SS GIs.

During the transfer process of T4SS GIs within the *C. difficile* population, the excision sites could become altered, resulting in different short-repeat sequences among T4SS GIs and different insertion sites in the chromosome. The new excision sites could potentially extend across the length of the GI and promote the integration of host genes located in the flank regions of the T4SS GI (Fig 3). The integrated host genes could then be transferred to recipient cells with T4SS GIs and cause the formation of the highly mosaic genomes characteristic of *C. difficile*. Gene exchange and new gene acquisition could be repeated multiple times during the transfer process among bacterial cells.

This study represent the first time that T4SS-type GIs were identified in *C. difficile*, and they were defined as a special type of CTn that mediated transference of genetic materials between host and recipient bacterial cells. This type of CTn has the following characteristics: (1) type-IVC secretion system genes (*VirB4*, *VirB6*, and *VirD4*) are located and clustered in these GIs; (2) short, direct-repeat sequences are located at both ends of the GIs; (3) multiple mobile element-related genes such as integrase, *Xis*, and *mobA* can be found in the GIs; (4) the T4SS region of GIs usually show very high similarity, although the gene contents and GI lengths are quite variable.

In this study, we employed bioinformatics and PCR methods to show that T4SS genes exist widely in *C. difficile*, revealing a novel way in which *C. difficile* can transfer DNA elements among strains, including resistance and virulence genes. The function of proteins such as *VirB4*, *VirB6*, and *VirD4* in transporting DNA in *S. suis* has been demonstrated by constructing strains with each individual gene knocked out (e.g. $\Delta virB4-89K$ and $\Delta virD4-89K$) (Li et al. 2011). These knockout

organisms were significantly deficient in transconjugation (Li et al. 2011). Similar to their homologous counterparts in *S. suis*, the *VirB4* *VirB6*, and *VirD4* genes in *C. difficile* may be involved in similar molecular mechanisms important for the genetic exchange of *C. difficile*. In prospective studies, we plan to knock out the *VirB4* *VirB6*, and *VirD4* genes in *C. difficile* to investigate their exact functions in gene transfer.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 81301402) and 863 Project Nos. 2014AA021505, 2013ZX10004221, and 2013ZX10004-101-002.

Draft

References

- Brouwer, M.S., Roberts, A.P., Mullany, P., and Allan, E. 2012. In silico analysis of sequenced strains of *Clostridium difficile* reveals a related set of conjugative transposons carrying a variety of accessory genes. *Mob. Genet. Elements*, **2**(1): 8–12. doi: 10.4161/mge.19297.
- Brouwer, M.S., Warburton, P.J., Roberts, A.P., Mullany, P., and Allan, E. 2011. Genetic organisation, mobility and predicted functions of genes on integrated, mobile genetic elements in sequenced strains of *Clostridium difficile*. *PLoS One*, **6**(8): e23014. doi: 10.1371/journal.pone.0023014.
- Brouwer, M.S.M., Roberts, A.P., Hussain, H., Williams, R.J., Allan, E., and Mullany, P. 2013. Horizontal gene transfer converts non-toxigenic *Clostridium difficile* strains into toxin producers. *Nat. Commun.* **4**(10): 2601–2601. doi: 10.1038/ncomms3601.
- Chen, C., Tang, J., Dong, W., Wang, C., Feng, Y., Wang, J., Zheng, F., Pan, X., Liu, D., Li, M., Song, Y., Zhu, X., Sun, H., Feng, T., Guo, Z., Ju, A., Ge, J., Dong, Y., Sun, W., Jiang, Y., Wang, J., Yan, J., Yang, H., Wang, X., Gao, G.F., Yang, R., Wang, J., and Yu, J. 2007. A glimpse of streptococcal toxic shock syndrome from comparative genomics of *S. suis* 2 Chinese isolates. *PLoS One*, **2**(3): e315. doi: 10.1371/journal.pone.0000315.
- Chen, C., Zhang, W., Zheng, H., Lan, R., Wang, H., Du, P., Bai, X., Ji, S., Meng, Q., Jin, D, Liu, K., Jing, H., Ye, C., Gao, G.F., Wang, L., Gottschalk, M., and Xu, J. 2013. Minimum core genome sequence typing of bacterial pathogens: a unified approach for clinical and public health microbiology. *J. Clin. Microbiol.*, **51**(8): 2582–2591. doi: 10.1128/JCM.00535-13.
- Collins DA, Hawkey PM, and Riley TV. 2013. Epidemiology of *Clostridium difficile* infection in Asia. *Antimicrob. Resist. Infect. Control*, **2**(1): 21. doi: 10.1186/2047-2994-2-21.
- Delcher, A.L., Bratke, K.A., Powers, E.C., and Salzberg, S.L. 2007. Identifying bacterial genes and

- endosymbiont DNA with Glimmer. *Bioinformatics*, **23**(6): 673–679. doi: 10.1093/bioinformatics/btm009.
- Dexi, B., Linmeng, L., Cui, T., Zixin, D., Kumar, R., and Hong-Yu, O. 2012. SecReT4: A web-based bacterial type IV secretion system resource. *Nucleic Acids Res.*, **41**(Database issue): D660–D665. doi: 10.1093/nar/gks1248.
- Drudy, D., Gerding, D.N., Stabler, R.A., Brazier, J.S., Wren, B.W., Hinds, J., Trinh, H.T., Songer, J.G., Witney, A.A, Hinds, J., and Wren, B.W. 2006. Comparative phylogenomics of *Clostridium difficile* reveals clade specificity and microevolution of hypervirulent strains. *J Bacteriol.*, **188**(20): 7297–7305. doi: 10.1128/JB.00664-06.
- Du, P., Cao, B., Wang, J., Li, W., Jia, H., Zhang, W., Lu, J., Li, Z., Yu, H., Chen, C., and Cheng, Y. 2014. Sequence variation in *tcdA* and *tcdB* of *Clostridium difficile*: ST37 with truncated. *J. Clin. Microbiol.* **52**(9): 3264–3270. doi: 10.1128/JCM.03487-13.
- Griffiths, D., Fawley, W., Kachrimanidou, M., Bowden, R., Crook, D.W., Fung, R., Golubchik, T., Harding, R.M., Jeffery, K.J., Jolley, K.A., Kirton, R., Peto, T.E., Rees, G., Stoesser, N., Vaughan, A., Walker, A.S., Young, B.C., Wilcox, M., and Dingle, K.E. 2009. Multilocus sequence typing of *Clostridium difficile*. *J Clin. Microbiol.*, **48**(3): 770–778. doi: 10.1128/JCM.01796-09.
- He, M., Sebaihia, M., Lawley, T.D., Stabler, R.A., Dawson, L.F., Martin, M.J., Holt, K.E., Seth-Smith, H.M.B., Quail, M.A., Rance, R., Brooks, K., Churcher, C., Harris, D., Bentley, S.D., Burrows, C., Clark, L., Corton, C., Murray, V., Rose, G., Thurston, S., van Tonder, A., Walker, D., Wren, B.W., Dougan, G., and Parkhill, J. 2010. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc. Natl. Acad. Sci. U. S. A.*, **107**(16): 7527–7532. doi:

10.1073/pnas.0914322107.

Huang, H., Weintraub, A., Fang, H., and Nord, C.E. 2009. Antimicrobial resistance in *Clostridium difficile*. *Int. J. Antimicrob. Agents.*, **34**(6): 516–522. doi: 10.1016/j.ijantimicag.2009.09.012.

Kato, H., Kato, N., Katow, S., and Maegawa, T., Nakamura, S., and Lysterly, D.M. 1999. Deletions in the repeating sequences of the toxin A gene of toxin A-negative, toxin B-positive *Clostridium difficile* strains. *FEMS Microbiol. Lett.* **175**(2): 197–203. doi: 10.1111/j.1574-6968.1999.tb13620.x.

Kurtz, S., Phillippy, A., Delcher, A., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S. 2004. Versatile and open software for comparing large genomes. *Genome Biology* **5**(2). R12. doi: 10.1186/gb-2004-5-2-r12.

Li, L., Stoeckert, C.J., and Roos, D.S. 2003. OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res.*, **13**(9): 2178–2189. doi: 10.1101/gr.1224503

Li, M., Shen, X., Yan, J., Han, H., Zheng, B., Liu, D., Cheng, H., Zhao, Y., Rao, X., Wang, C., Tang, J., Hu, F., and Gao, G.F. 2011. GI-type T4SS-mediated horizontal transfer of the 89K pathogenicity island in epidemic *Streptococcus suis* serotype 2. *Mol. Microbiol.*, **79**(6): 1670–1683. doi: 10.1111/j.1365-2958.2011.07553.x.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. 2010. De novo assembly of human genomes with massively parallel short read sequencing. *Genome Res.*, **20**(2): 265–272. doi: 10.1101/gr.097261.109.

Loo, V.G., Poirier, L., Miller, M.A., Oughton, M., Libman, M.D., Michaud, S., Bourgault, A.M., Nguyen, T., Frenette, C., Kelly, M., Vibien, A., Brassard, P., Fenn, S., Dewar, K., Hudson, T.J., Horn, R., René, P., Monczak, Y., and Dascal, A. 2006. A predominantly clonal

- multi-institutional outbreak of *Clostridium difficile*-Associated diarrhea with high morbidity and mortality. *N. Engl. J. Med.*, **353**(23): 2442–2449. doi: 10.1056/NEJMoa051639.
- O'Neillf, G.L., Ogunsola, F.T., and Duerden, J.S.B.I. 1996. Modification of a PCR ribotyping method for application as a routine typing scheme for *Clostridium difficile*. *Anaerobe* **2**(4): 205–209. doi: 10.1006/anae.1996.0028
- Sam, M.D., Cascio, D., Johnson, R.C., and Clubb, R.T. 2004. Crystal structure of the excisionase–DNA complex from bacteriophage lambda. *J. Mol. Biol.*, **338**(2): 229–240. doi: 10.1016/j.jmb.2004.02.053.
- Schwan, C., Stecher, B., Tzivelekidis, T., van Ham, M., Rohde, M., Hardt, W.D., Wehland, J., Aktories, K. 2009. *Clostridium difficile* toxin CDT induces formation of microtubule-based protrusions and increases adherence of bacteria. *PLoS Pathog.*, **5**(10): e1000626. doi: 10.1371/journal.ppat.1000626.
- Sebaihia, M., Wren, B.W., Mullany, P., Fairweather, N.F., Minton, N., Stabler, R., Thomson, N.R., Roberts, A.P., Cerdeño-Tárraga, A.M., Wang, H., Holden, M.T., Wright, A., Churcher, C., Quail, M.A., Baker, S., Bason, N., Brooks, K., Chillingworth, T., Cronin, A., Davis, P., Dowd, L., Fraser, A., Feltwell, T., Hance, Z., Holroyd, S., Jagels, K., Moule, S., Mungall, K., Price, C., Rabinowitsch, E., Sharp, S., Simmonds, M., Stevens, K., Unwin, L., Whithead, S., Dupuy, B., Dougan, G., Barrell, B., and Parkhill, J. 2006. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat. Genet.*, **38**(7): 779–786.
- Seiter, C. 1992. Sequencher 2.0. *Macworld*, **9**(12): 274.
- Tamura, K., Dudley, J., Nei, M., and Kumar, S. 2007. MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) Software Version 4.0. *Mol. Biol. Evol.*, **24**(8): 1596–1599. doi:

10.1093/molbev/msm092.

- Warny, M., Pepin, J., Fang, A., Killgore, G., Thompson, A., Brazier, J., Frost, E., and McDonald, L.C. 2005. Toxin production by an emerging strain of *Clostridium difficile* associated with outbreaks of severe disease in North America and Europe. *Lancet*, **366**(9491): 1079–1084. doi: 10.1016/S0140-6736(05)67420-X.
- Zhang W, Yu WW, Liu D, Li M, DU PC, Wu YL, Gao GF, Chen C. 2013. T4SP: A novel tool and database for type IV secretion systems in bacterial genomes. *Biomed. Environ. Sci.*, **26**(7): 614–617. doi: 10.3967/0895-3988.2013.07.015.
- Zhang, W., Du, P., Zheng, H., Yu, W., Wan, L., and Chen, C. 2014. Whole-genome sequence comparison as a method for improving bacterial species definition. *J. Gen. Appl. Microbiol.*, **60**(2): 75–78. doi: 10.2323/jgam.60.75.
- Zhang, W., Rong, C., Chen, C., and Gao, G.F. 2012. Type-IVC secretion system: a novel subclass of type IV secretion system (T4SS) common existing in gram-positive genus *Streptococcus*. *PLoS One* **7**(10): e46390. doi: 10.1371/journal.pone.0046390.

Table 1. List of 10 T4SS GIs identified in *C. difficile* strains. “Left site” and “Right site” represent the GI start and end site in the corresponding chromosome, while “Insertion site on BI1” represent the position of GI in *C. difficile* BI1.

Strain	T4SS GI	Insertion site in BI1	Left site	Right site	GI length	Repeat	CTn name
<i>C. difficile</i> 630	GI1	466,656	480,392	519,797	39,406	CACAT/CACAT	CTn2
<i>C. difficile</i> 630	GI2	1,175,251	1,284,321	1,314,877	30,557	AATTTA/AATTT A	CTn4
<i>C. difficile</i> 630	GI3	2,052,815	2,137,462	2,183,040	45,579	GTTGA/GTTGA	CTn5
<i>C. difficile</i> 2007855	GI4	3,760,138	3,771,587	3,821,434	49,848	GTTC/GTCTC	CTn5-li ke
<i>C. difficile</i> ATCC43255	GI5	3,760,138	3,610,480	3,678,086	67,607	GTTC/GTCTC	New
<i>C. difficile</i> CF5	GI6	466,669	430,059	479,978	49,920	TGAGACGGTA G/TGAGACTGT AG	CTn5-li ke
<i>C. difficile</i> M68	GI7	466,669	407,642	457,561	49,920	TGAGACGGTA G/TGAGACTGT AG	CTn5-li ke
<i>C. difficile</i> M120	GI8	466,600	418,467	547,658	129,192	GAGAT/GAGAT	Tn6164
<i>C. difficile</i> BJ08	GI9	466,669	342,674	422,543	79,870	TGAGACGGTA	New

						G/TGAGACTGT AG	
<i>C. difficile</i> <i>R20291</i>	GI10	2,052,815	2,040,400	2,125,358	84,959	GTTGA/GTTGA	Tn6103

Draft

Figure legends

Fig 1. A) Phylogenetic tree of 10 *C. difficile* strains based on 66,192 core SNPs and the genome locations of 10 T4SS GIs in *C. difficile* strain BI1. B) Detection of VirD4 homologs in 24 *C. difficile* strains by PCR analysis using the primers 5'-TGCAAGATAAGGCAAAGTTTC-3' and 5'-ACTTCTGAAGCGTCTATCATATC-3'

Fig 2. Phylogenetic tree of T4SS genes in GIs and co-lineage comparisons of 10 T4SS GIs. The left tree is a neighbor-joining tree based on 3 genes (*VirB4*, *VirB6* and *VirD4*) in 10 T4SS GIs. The start and end positions of each GI are represented by the left and right ends of black lines with arrows. Genes with various functions are presented using arrows with different colors. The grey/black lines between the GIs represent GIs with similar DNA sequences.

Fig 3. Schematic representation of changes in the short repeat sequences between GI3 and GI6, which caused a 2-bp deletion at both the 5' end of the GI and at the 3' end (located 5.8 kb downstream). The pink cylinder represents the repeat sequence "GTTGA," whereas the green cylinder represents the sequence "TGAGACGGTAG/TGAGACTGTAG." The nucleoside sequences of the short repeat sequences are marked by dotted squares. Red lines represent the GI3 region, which was also found in GI6. Five genes (CD630_18650–CD630_18690) of GI3 (yellow cylinder) were replaced by 2 genes (CF5GL000410 and CF5GL000411) in T4SS GI6 (blue cylinder).

Supporting information captions

Table S1. Detailed information for the 10 *C. difficile* genomes analyzed in this study

Table S2. Detailed information for the 24 *C. difficile* strains used in PCR experiments

Table S3. Average genome identity (ANI) values of pairs of 10 different *C. difficile* genomes

Table S4. Annotation information for the genes in 10 T4SS GIs and their candidate sources

Table S5. Functional categories of genes in 10 T4SS GIs. Red represents the presence of a given gene, while white represents its absence. The numbers shown represent the gene annotation numbers for a given function, which were determined by comparing information from the NT, NR, COG, and KEGG databases.

Fig S1. GC-content skew at the genome level in 10 *C. difficile* strains

Fig S2. Phylogenetic tree generated based on the topoisomerase IA gene

Fig S3. Gene sources among 10 T4SS GIs at the genus level

Fig 1. A) Phylogenetic tree of 10 *C. difficile* strains based on 66,192 core SNPs and the genome locations of 10 T4SS GIs in *C. difficile* strain BI1. B) Detection of VirD4 homologs in 24 *C. difficile* strains by PCR analysis using the primers 5'-TGCAAGATAAGGCAAAGTTTC-3' and 5'-ACTTCTGAAGCGTCTATCATATC-3'

Figure 1

317x301mm (300 x 300 DPI)

Fig 2. Phylogenetic tree of T4SS genes in GIs and co-lineage comparisons of 10 T4SS GIs. The left tree is a neighbor-joining tree based on 3 genes (VirB4, VirB6 and VirD4) in 10 T4SS GIs. The start and end positions of each GI are represented by the left and right ends of black lines with arrows. Genes with various functions are presented using arrows with different colors. The grey/black lines between the GIs represent GIs with similar DNA sequences.

Figure 2
567x388mm (150 x 150 DPI)

Fig 3. Schematic representation of changes in the short repeat sequences between GI3 and GI6, which caused a 2-bp deletion at the 5' end of the GI and at the 3' end (located 5.8 kb downstream). The pink cylinder represents the repeat sequence "GTTGA," whereas the green cylinder represents "TGAGACGGTAG/TGAGACTGTAG." The nucleoside sequences of short repeat sequences are marked by dotted squares. Red lines represent the GI3 region, which was also found in GI6. Five genes (CD630_18650-CD630_18690) of GI3 (yellow cylinder) were replaced by two genes (CF5GL000410 and CF5GL000411) in T4SS GI6 (blue cylinder).

Figure 3

60x28mm (300 x 300 DPI)

Supplemental Table 1. Detailed information of 10 *C. difficile* strains from NCBI.

Strains	Genotype	Year collection	Source	Site collection	Sequence type	GenBank Accession no.
ATCC43255	A+B+	2001	Human	USA	3	NC_013974
630	A+B+	1982	Human	Switzerland	54	NC_009089
R20291	A+B+	2006	Human	UK	1	NC_013316
CD196	A+B+	1985	Human	France	1	NC_013315
M120	A+B+	2007	Human	UK	11	NC_017174
2007855	A+B+	2007	Bovine	USA	1	NC_017178
BI1	A+B+	1988	Human	USA	1	NC_017177
BJ08	A-B+	2010	outpatient	BJ	37	CP003939
M68	A-B+	2006	Human	Ireland	37	NC_017175
CF5	A-B+	1995	Human	Belgium	37	NC_017173

Supplemental Table 2. Detailed information of *C. difficile* strains used in PCR experiments.

Strains	Genotype	Year collection	Source	Site Collection*	Sequence type
UK1	A+B+	unknown	unknown	UK	1
GZ5	A+B+	1980' s	inpatient	GZ	2
ATCC 9689	A+B+	unknown	unknown	ATCC	3
ZR17	A+B+	2010	inpatient	BJ	5
ZR75	A+B+	2010	inpatient	BJ	8
GZ1	A+B+	1980' s	inpatient	GZ	35
VPI10463	A+B+	unknown	unknown	Japan	46
ZR50	A+B+	2010	inpatient	BJ	53
ZR4	A+B+	2010	outpatient	BJ	54
ZR 5	A+B+	2010	outpatient	BJ	55
ZR41	A+B+	2010	inpatient	BJ	92
ZR27	A+B+	2010	outpatient	BJ	99
ZR 2	A+B+	2010	outpatient	BJ	102
ZR77	A+B+	2010	inpatient	BJ	129
US1	A-B+	unknown	unknown	US	37
BJ08	A-B+	2010	outpatient	BJ	37

GZ2	A-B+	1980' s	inpatient	GZ	37
ZR12	A-B+	2010	outpatient	BJ	15
ZR34	A-B+	2010	inpatient	BJ	48
GZ15	A-B+	1980' s	inpatient	GZ	119
ZR10	A-B+	2010	outpatient	BJ	118
ZR49	A-B+	2010	outpatient	BJ	100
ZR25	A-B+	2010	outpatient	BJ	117
ZR20	A+B+	2010	inpatient	BJ	54
ZR30	A+B+	2010	outpatient	BJ	54
ZR74	A+B+	2011	inpatient	BJ	54

*: "BJ" represents the Beijing city in China; "GZ" represents Guangzhou city in China

Supplemental Table 3. Average genome identity (ANI) value of pairs of 10 *C. difficile* genomes.

ANI	<i>2007855</i>	<i>ATCC43255</i>	<i>BJ08</i>	<i>CF5</i>	<i>M120</i>	<i>M68</i>	<i>630</i>	<i>CD196</i>	<i>R20291</i>	<i>BI1</i>
<i>C. difficile 2007855</i>	100.00 %	98.18 %	97.44 %	97.81 %	95.80 %	97.58 %	98.32 %	99.89 %	99.85 %	99.89 %
<i>C. difficile ATCC43255</i>	98.22 %	100.00 %	97.37 %	97.77 %	95.77 %	97.54 %	98.43 %	98.27 %	98.20 %	98.27 %
<i>C. difficile BJ08</i>	97.39 %	97.39 %	100.00 %	99.59 %	95.79 %	99.82 %	97.61 %	97.48 %	97.38 %	97.47 %
<i>C. difficile CF5</i>	97.76 %	97.71 %	99.54 %	100.00 %	95.88 %	99.64 %	97.95 %	97.81 %	97.76 %	97.80 %
<i>C. difficile M120</i>	96.04 %	95.95 %	95.99 %	96.13 %	100.00 %	96.05 %	96.06 %	96.14 %	96.07 %	96.14 %
<i>C. difficile M68</i>	97.48 %	97.40 %	99.74 %	99.56 %	95.75 %	100.00 %	97.61 %	97.43 %	97.38 %	97.42 %
<i>C. difficile 630</i>	98.12 %	98.11 %	97.36 %	97.87 %	95.53 %	97.44 %	100.00 %	98.19 %	98.09 %	98.19 %
<i>C. difficile CD196</i>	99.96 %	98.31 %	97.62 %	97.90 %	95.97 %	97.66 %	98.42 %	100.00 %	99.97 %	99.97 %
<i>C. difficile R20291</i>	99.90 %	98.22 %	97.49 %	97.83 %	95.88 %	97.59 %	98.37 %	99.95 %	100.00 %	99.95 %
<i>C. difficile BI1</i>	99.97 %	98.29 %	97.60 %	97.88 %	95.95 %	97.64 %	98.40 %	99.98 %	99.98 %	100.00 %

Supplemental Table 4. Annotation information of Genes in 10 T4SS

GIs and their candidate source.

GI ID	Gene Name	Gene Annotation	Gene Source
GI1	CD630_04080		<i>Streptococcus pyogenes</i> ICESp2905 DNA containing erm(TR)-carrying element and tet(0) fragment, strain iB21
GI1	CD630_04081	membrane protein	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> RE378 DNA
GI1	CD630_04090	replication initiation protein	<i>Streptococcus intermedius</i> B196
GI1	CD630_04100	DNA replication protein	<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C1050
GI1	CD630_04110		<i>Streptococcus equi</i> subsp. <i>equi</i> 4047
GI1	CD630_04120	Type IV secretory pathway, VirD4 components	<i>Streptococcus intermedius</i> C270
GI1	CD630_04121		
GI1	CD630_04130	single-strand binding protein	<i>Streptococcus intermedius</i> B196
GI1	CD630_04140	conjugative transposon membrane protein	<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C1050
GI1	CD630_04150	membrane protein	<i>Streptococcus anginosus</i> C238
GI1	CD630_04160	exported protein	<i>Schistosoma mansoni</i> hypothetical protein (Smp_090990) mRNA, complete cds
GI1	CD630_04170	membrane protein	<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C1050
GI1	CD630_04180	Type IV secretory pathway, VirB4 components	<i>Streptococcus intermedius</i> B196
GI1	CD630_04190		<i>Streptococcus intermedius</i> B196
GI1	CD630_04191		
GI1	CD630_04200	cell surface protein	<i>Streptococcus intermedius</i> B196
GI1	CD630_042	Topoisomerase IA	<i>Streptobacillus moniliformis</i> DSM 12112

	10		
GI1	CD630_042 20		<i>Streptobacillus moniliformis</i> DSM 12112
GI1	CD630_042 30	DNA methylase	<i>Streptococcus intermedius</i> B196
GI1	CD630_042 40	single-strand binding protein	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> RE378 DNA
GI1	CD630_042 50	mobilization protein	Clostridiales genomosp. BVAB3 str. UPII9-5
GI1	CD630_042 60	mobilization protein	<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i> H70
GI1	CD630_042 70	transcriptional regulators	<i>Slackia heliotrinireducens</i> DSM 20476
GI1	CD630_042 80	AraC-type DNA-binding domain-containin g proteins	Clostridiales genomosp. BVAB3 str. UPII9-5
GI1	CD630_042 90		<i>Streptococcus anginosus</i> C238
GI1	CD630_043 00	ABC-type cobalt transport system, permease component CbiQ and related transporters	<i>Streptococcus anginosus</i> C238
GI1	CD630_043 10	ABC-type cobalt transport system, ATPase component	<i>Streptococcus anginosus</i> C238
GI1	CD630_043 20	ABC-type multidrug transport system, ATPase and permease components	<i>Streptococcus anginosus</i> C238
GI1	CD630_043 30	ABC-type multidrug transport system, ATPase and permease components	<i>Streptococcus anginosus</i> C238
GI1	CD630_043 40	Na ⁺ -driven multidrug efflux pump	<i>Streptococcus pyogenes</i> ICESp2905 DNA containing erm(TR)-carrying element and tet(0) fragment, strain iB21

GI1	CD630_043 41		
GI1	CD630_043 50	sigma factor	Streptococcus pyogenes ICESp2905 DNA containing erm(TR)-carrying element and tet(0) fragment, strain iB21
GI1	CD630_043 51		
GI1	CD630_043 52		
GI1	CD630_043 60	Site-specific recombinases, DNA invertase Pin homologs	Streptococcus pyogenes ICESp2905 DNA containing erm(TR)-carrying element and tet(0) fragment, strain iB21
GI10	CDR20291_1741	membrane protein	Streptococcus equi subsp. zooepidemicus ATCC 35246
GI10	CDR20291_1742	conjugative transposon protein	Filifactor alocis ATCC 35896
GI10	CDR20291_1744	Site-specific recombinases, DNA invertase Pin homologs	Ruminococcus torques L2-14 draft genome
GI10	CDR20291_1745		Roseburia hominis A2-183
GI10	CDR20291_1746		Roseburia hominis A2-183
GI10	CDR20291_1747	helix-turn-helix protein	Roseburia hominis A2-183
GI10	CDR20291_1748	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Roseburia hominis A2-183
GI10	CDR20291_1749	Signal transduction histidine kinase	Uncultured organism clone 7 genomic sequence
GI10	CDR20291_1750	ABC-type multidrug transport system, ATPase	Ruminococcus obeum A2-162 draft genome

		component	
GI10	CDR20291_1751	lantibiotic ABC transporter permease	Uncultured organism clone 7 genomic sequence
GI10	CDR20291_1752	lantibiotic ABC transporter permease	Uncultured organism clone 7 genomic sequence
GI10	CDR20291_1753		Uncultured organism clone 7 genomic sequence
GI10	CDR20291_1754	rna polymerase	Roseburia hominis A2-183
GI10	CDR20291_1755	sigma-24 (FecI)	Roseburia intestinalis XB6B4 draft genome
GI10	CDR20291_1756	rna polymerase, sigma-24 subunit, ecf subfamily	Roseburia hominis A2-183
GI10	CDR20291_1757		Roseburia hominis A2-183
GI10	CDR20291_1759	toxin-antitoxin system, toxin component, RelE family	Roseburia hominis A2-183
GI10	CDR20291_1760	DNA-damage-inducible protein J	Roseburia hominis A2-183
GI10	CDR20291_1761		Coproccoccus sp. ART55/1 draft genome
GI10	CDR20291_1762	phage protein	Ruminococcus bromii L2-63 draft genome
GI10	CDR20291_1763	replicative dna helicase	Ruminococcus bromii L2-63 draft genome
GI10	CDR20291_1764		Faecalibacterium prausnitzii SL3/3 draft genome
GI10	CDR20291_1765		Clostridium saccharolyticum-like K10 draft genome
GI10	CDR20291_1766	transcriptional regulators	Clostridiales sp. SM4/1 draft genome
GI10	CDR20291_1767		Clostridiales sp. SM4/1 draft genome
GI10	CDR20291_1768	serine/arginine repetitive matrix protein 2	Clostridiales sp. SM4/1 draft genome
GI10	CDR20291_1769	DNA primase (bacterial type)	Clostridiales sp. SM4/1 draft genome

GI10	CDR20291_1770	P-loop ATPase and inactivated derivatives	Clostridium saccharolyticum-like K10 draft genome
GI10	CDR20291_1771	Site-specific recombinases, DNA invertase Pin homologs	Clostridium sp. SY8519 DNA
GI10	CDR20291_1772	Site-specific recombinases, DNA invertase Pin homologs	Uncultured bacterium EB2 genomic sequence
GI10	CDR20291_1773		Uncultured bacterium EB2 genomic sequence
GI10	CDR20291_1774	RNA methyltransferase	Uncultured bacterium EB2 genomic sequence
GI10	CDR20291_1775	Response regulator of the LytR/AlgR family	Bifidobacterium breve ACS-071-V-Sch8b
GI10	CDR20291_1776	single-strand binding protein	Streptococcus equi subsp. equi 4047
GI10	CDR20291_1777	TnpV	Ruminococcus bromii L2-63 draft genome
GI10	CDR20291_1778	Transcriptional regulators	Lactobacillus ruminis ATCC 27782
GI10	CDR20291_1779	Na ⁺ -driven multidrug efflux pump	Streptococcus pyogenes ICESp2905 DNA containing erm(TR)-carrying element and tet(0) fragment, strain iB21
GI10	CDR20291_1780		Faecalibacterium prausnitzii SL3/3 draft genome
GI10	CDR20291_1781	phage protein	Eubacterium siraeum 70/3 draft genome
GI10	CDR20291_1782		Faecalibacterium prausnitzii SL3/3 draft genome
GI10	CDR20291_1783		Clostridiales sp. SS3/4 draft genome
GI10	CDR20291_1784	ATP-dependent exoDNase (exonuclease V), alpha subunit - helicase superfamily I member	Faecalibacterium prausnitzii SL3/3 draft genome
GI10	CDR20291_1785	P-loop ATPase and	Clostridium saccharolyticum-like K10

	1786	inactivated derivatives	draft genome
GI10	CDR20291_1787		Clostridium saccharolyticum-like K10 draft genome
GI10	CDR20291_1788	Site-specific recombinases, DNA invertase Pin homologs	Clostridium saccharolyticum-like K10 draft genome
GI10	CDR20291_1789	conjugative transposon membrane protein	Streptococcus anginosus subsp. whileyi MAS624 DNA
GI10	CDR20291_1790	Type IV secretory pathway, VirB4 components	Streptococcus intermedius B196
GI10	CDR20291_1791		Streptococcus constellatus subsp. pharyngis C1050
GI10	CDR20291_1792	cell surface protein	Streptococcus intermedius B196
GI10	CDR20291_1793	Topoisomerase IA	Streptobacillus moniliformis DSM 12112
GI10	CDR20291_1794		Streptobacillus moniliformis DSM 12112
GI10	CDR20291_1795	DNA methylase	Streptococcus intermedius B196
GI10	CDR20291_1796		Streptococcus anginosus C238
GI10	CDR20291_1797	ATP-dependent endonuclease of the OLD family	Listeria monocytogenes strain SLCC2376, serotype 4c
GI10	CDR20291_1798	conjugative transposon protein	Finegoldia magna ATCC 29328 DNA
GI10	CDR20291_1799		Streptococcus gallolyticus subsp. gallolyticus ATCC BAA-2069 complete chromosome sequence, strain ATCC BAA-2069
GI10	CDR20291_1800	conjugative transposon mobilization protein	Streptococcus anginosus C238
GI10	CDR20291_1801	exported protein	Clostridiales sp. SSC/2 draft genome
GI10	CDR20291_1802	Polyferredoxin	Clostridiales sp. SSC/2 draft genome

GI10	CDR20291_1803	ABC-type antimicrobial peptide transport system, permease component	Clostridiales sp. SSC/2 draft genome
GI10	CDR20291_1804	ABC-type antimicrobial peptide transport system, ATPase component	Clostridiales sp. SSC/2 draft genome
GI10	CDR20291_1805	ABC-type antimicrobial peptide transport system, permease component	Enterococcus faecalis 62
GI10	CDR20291_1806	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Enterococcus faecium Aus0085 plasmid p1, complete sequence
GI10	CDR20291_1807	Signal transduction histidine kinase	Enterococcus faecalis 62
GI10	CDR20291_1808	sigma factor	Streptococcus intermedius C270
GI2	CD630_10910	Integrase	Ruminococcus torques L2-14 draft genome
GI2	CD630_10920	xis; excisionase	Ruminococcus torques L2-14 draft genome
GI2	CD630_10921	conjugative transposon protein	Ruminococcus torques L2-14 draft genome
GI2	CD630_10940	conjugative transposon protein	Ruminococcus torques L2-14 draft genome
GI2	CD630_10950	lantibiotic ABC transporter	Clostridium saccharolyticum-like K10 draft genome

		permease	
GI2	CD630_10960	lantibiotic ABC transporter permease	Clostridium saccharolyticum-like K10 draft genome
GI2	CD630_10970	ABC-type multidrug transport system, ATPase component	Roseburia hominis A2-183
GI2	CD630_10980	Signal transduction histidine kinase	Uncultured organism clone 20 genomic sequence
GI2	CD630_10990	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Ruminococcus obeum A2-162 draft genome
GI2	CD630_10991		Uncultured organism clone VC1DB32TF genomic sequence
GI2	CD630_11000	conjugative transposon protein	Ruminococcus obeum A2-162 draft genome
GI2	CD630_11010	mobilization protein	Ruminococcus torques L2-14 draft genome
GI2	CD630_11020	Type IV secretory pathway, VirD2 components (relaxase)	Ruminococcus torques L2-14 draft genome
GI2	CD630_11030		Uncultured organism clone VC1A912TR genomic sequence
GI2	CD630_11040		Uncultured organism clone 7 genomic sequence
GI2	CD630_11041		Clostridium saccharolyticum-like K10 draft genome
GI2	CD630_11042		Ruminococcus torques L2-14 draft genome
GI2	CD630_11050	DNA primase	Ruminococcus torques L2-14 draft genome
GI2	CD630_11060	Topoisomerase IA	Ruminococcus torques L2-14 draft genome

GI2	CD630_11061		Ruminococcus bromii L2-63 draft genome
GI2	CD630_11070	membrane protein	Ruminococcus torques L2-14 draft genome
GI2	CD630_11071		Ruminococcus torques L2-14 draft genome
GI2	CD630_11080	DNA-repair protein	Ruminococcus torques L2-14 draft genome
GI2	CD630_11090	DNA modification methylase	Ruminococcus torques L2-14 draft genome
GI2	CD630_11100	Type IV secretory pathway, VirB4 components	Ruminococcus torques L2-14 draft genome
GI2	CD630_11110	membrane protein	Ruminococcus torques L2-14 draft genome
GI2	CD630_11120	membrane protein	Ruminococcus torques L2-14 draft genome
GI2	CD630_11130	conjugative transfer protein	Clostridium saccharolyticum WM1
GI2	CD630_11150	Type IV secretory pathway, VirD4 components	Ruminococcus torques L2-14 draft genome
GI2	CD630_11160		Coproccoccus sp. ART55/1 draft genome
GI2	CD630_11170		Ruminococcus torques L2-14 draft genome
GI2	CD630_11180		Ruminococcus torques L2-14 draft genome
GI3	CD630_18450	membrane protein	Streptococcus equi subsp. zooepidemicus ATCC 35246
GI3	CD630_18460		Streptococcus agalactiae A909
GI3	CD630_18470		Filifactor alocis ATCC 35896
GI3	CD630_18480		Streptococcus anginosus C238
GI3	CD630_18490	Type IV secretory pathway, VirD4 components	Streptococcus intermedius C270
GI3	CD630_18500	AraC family transcription regulator	Streptococcus parasanguinis FW213
GI3	CD630_18510	single-stranded DNA binding	Streptococcus anginosus C238

		protein	
GI3	CD630_18520	conjugative transposon membrane protein	Streptococcus anginosus C238
GI3	CD630_18530	conjugative transposon membrane protein	Streptococcus anginosus subsp. whileyi MAS624 DNA
GI3	CD630_18540	conjugative transposon membrane exported protein	Schistosoma mansoni hypothetical protein (Smp_090990) mRNA, complete cds
GI3	CD630_18550	conjugative transposon membrane protein	Streptococcus pyogenes ICESp2905 DNA containing erm(TR)-carrying element and tet(0) fragment, strain iB21
GI3	CD630_18560	Type IV secretory pathway, VirB4 components	Streptococcus anginosus C238
GI3	CD630_18570	Cell wall-associated hydrolases (invasion-associated proteins)	Streptococcus intermedius B196
GI3	CD630_18571		
GI3	CD630_18580	cell surface protein	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI3	CD630_18590	Topoisomerase IA	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI3	CD630_18591	conjugative transposon regulatory protein	Streptococcus equi subsp. zooepidemicus H70
GI3	CD630_18600		Streptococcus anginosus C238
GI3	CD630_18610	O-Methyltransferase involved in polyketide biosynthesis	Treponema denticola ATCC 35405
GI3	CD630_18620	DNA methylase	Streptococcus anginosus C238
GI3	CD630_18630		Streptococcus anginosus C238
GI3	CD630_18640	transcriptional regulator	Campylobacter hominis ATCC BAA-381

GI3	CD630_18650	ATP-dependent endonuclease of the OLD family	<i>Listeria monocytogenes</i> strain SLCC2376, serotype 4c
GI3	CD630_18660		<i>Finegoldia magna</i> ATCC 29328 DNA
GI3	CD630_18670		<i>Streptococcus gallolyticus</i> subsp. <i>gallolyticus</i> ATCC BAA-2069 complete chromosome sequence, strain ATCC BAA-2069
GI3	CD630_18680		<i>Streptococcus gallolyticus</i> subsp. <i>gallolyticus</i> ATCC BAA-2069 complete chromosome sequence, strain ATCC BAA-2069
GI3	CD630_18690	conjugative transposon mobilization protein	<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C1050
GI3	CD630_18700	conjugative transposon mobilization protein	<i>Enterococcus faecalis</i> 62
GI3	CD630_18710		<i>Clostridiales</i> sp. SSC/2 draft genome
GI3	CD630_18711		<i>Clostridiales</i> sp. SSC/2 draft genome
GI3	CD630_18720	Polyferredoxin	<i>Clostridiales</i> sp. SSC/2 draft genome
GI3	CD630_18730	ABC-type antimicrobial peptide transport system, permease component	<i>Clostridiales</i> sp. SSC/2 draft genome
GI3	CD630_18740	ABC-type antimicrobial peptide transport system, ATPase component	<i>Clostridiales</i> sp. SSC/2 draft genome
GI3	CD630_18750	ABC-type antimicrobial peptide transport system, permease	<i>Enterococcus faecalis</i> 62

		component	
GI3	CD630_18760	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Enterococcus faecium Aus0085 plasmid p1, complete sequence
GI3	CD630_18770	Signal transduction histidine kinase	Enterococcus faecalis 62
GI3	CD630_18780	sigma factor	Streptococcus intermedius C270
GI3	CD630_18782		Streptococcus constellatus subsp. pharyngis C1050
GI4	2007855GL003379		Streptococcus constellatus subsp. pharyngis C1050
GI4	2007855GL003380	sigma factor	Streptococcus intermedius C270
GI4	2007855GL003381		Streptococcus equi subsp. zooepidemicus H70
GI4	2007855GL003382	Signal transduction histidine kinase	Enterococcus faecalis 62
GI4	2007855GL003383	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Enterococcus faecium Aus0085 plasmid p1, complete sequence
GI4	2007855GL003384	ABC-type antimicrobial peptide transport system, permease component	Enterococcus faecalis 62
GI4	2007855GL003385	ABC-type antimicrobial peptide	Clostridiales sp. SSC/2 draft genome

		transport system, ATPase component	
GI4	2007855GL003386	ABC-type antimicrobial peptide transport system, permease component	Clostridiales sp. SSC/2 draft genome
GI4	2007855GL003387	Polyferredoxin	Clostridiales sp. SSC/2 draft genome
GI4	2007855GL003388		Clostridiales sp. SSC/2 draft genome
GI4	2007855GL003389	conjugative transposon mobilization protein	Streptococcus anginosus C238
GI4	2007855GL003390		Streptococcus gallolyticus subsp. gallolyticus ATCC BAA-2069 complete chromosome sequence, strain ATCC BAA-2069
GI4	2007855GL003391		Streptococcus gallolyticus subsp. gallolyticus ATCC BAA-2069 complete chromosome sequence, strain ATCC BAA-2069
GI4	2007855GL003392		Finegoldia magna ATCC 29328 DNA
GI4	2007855GL003393	ATP-dependent endonuclease of the OLD family	Listeria monocytogenes strain SLCC2376, serotype 4c
GI4	2007855GL003394	transcriptional regulator	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI4	2007855GL003395		Streptococcus anginosus C238
GI4	2007855GL003396	DNA methylase	Streptococcus intermedius B196
GI4	2007855GL003397		Streptobacillus moniliformis DSM 12112
GI4	2007855GL003398	Topoisomerase IA	Streptobacillus moniliformis DSM 12112
GI4	2007855GL003399	cell surface protein	Streptococcus intermedius B196
GI4	2007855GL003400		

GI4	2007855GL 003401		<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C1050
GI4	2007855GL 003402	Type IV secretory pathway, VirB4 components	<i>Streptococcus intermedius</i> B196
GI4	2007855GL 003403		<i>Schistosoma mansoni</i> hypothetical protein (Smp_090990) mRNA, complete cds
GI4	2007855GL 003404	conjugative transposon membrane protein	<i>Streptococcus anginosus</i> subsp. <i>whileyi</i> MAS624 DNA
GI4	2007855GL 003405	conjugative transposon membrane protein	<i>Streptococcus equi</i> subsp. <i>equi</i> 4047
GI4	2007855GL 003406	single-strand binding protein	<i>Streptococcus intermedius</i> B196
GI4	2007855GL 003407	Type IV secretory pathway, VirD4 components	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> AC-2713
GI4	2007855GL 003408	Acetyltransferases, including N-acetylases of ribosomal proteins	Uncultured bacterium EB5 genomic sequence
GI4	2007855GL 003409	aminoglycoside phosphotransferase	<i>Enterococcus faecium</i> aminoglycoside phosphotransferase (aph(2')-Ib) gene, complete cds
GI4	2007855GL 003410	P-loop ATPase and inactivated derivatives	<i>Eubacterium rectale</i> ATCC 33656
GI4	2007855GL 003411	DNA primase (bacterial type)	<i>Coproccoccus</i> sp. ART55/1 draft genome
GI4	2007855GL 003412		<i>Coproccoccus</i> sp. ART55/1 draft genome
GI4	2007855GL 003413	Site-specific recombinases, DNA invertase Pin homologs	<i>Eubacterium rectale</i> ATCC 33656
GI4	2007855GL 003414	Type IV secretory pathway, VirD4 components	<i>Streptococcus anginosus</i> C238
GI4	2007855GL 003415	conjugative transposon protein	<i>Streptococcus anginosus</i> C238
GI4	2007855GL	conjugative	<i>Filifactor alocis</i> ATCC 35896

	003416	transposon protein	
GI4	2007855GL003417	membrane protein	<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i> ATCC 35246
GI4	2007855GL003418	SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase	<i>Filifactor alocis</i> ATCC 35896
GI5	ATCC43255GL003442		<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C1050
GI5	ATCC43255GL003443	sigma factor	<i>Streptococcus intermedius</i> C270
GI5	ATCC43255GL003444		<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i> H70
GI5	ATCC43255GL003445	Signal transduction histidine kinase	<i>Enterococcus faecalis</i> 62
GI5	ATCC43255GL003446	vncS; sensor protein	
GI5	ATCC43255GL003447	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	<i>Enterococcus faecium</i> Aus0085 plasmid p1, complete sequence
GI5	ATCC43255GL003448	ABC-type antimicrobial peptide transport system, permease component	<i>Enterococcus faecalis</i> 62
GI5	ATCC43255GL003449	ABC-type antimicrobial peptide transport system, ATPase component	<i>Clostridiales</i> sp. SSC/2 draft genome
GI5	ATCC43255GL003450	ABC-type antimicrobial	<i>Clostridiales</i> sp. SSC/2 draft genome

		peptide transport system, permease component	
GI5	ATCC43255 GL003451	Polyferredoxin	Clostridiales sp. SSC/2 draft genome
GI5	ATCC43255 GL003452		Clostridiales sp. SSC/2 draft genome
GI5	ATCC43255 GL003453		Clostridiales sp. SSC/2 draft genome
GI5	ATCC43255 GL003454		Enterococcus faecium D0 plasmid 3, complete sequence
GI5	ATCC43255 GL003455		Enterococcus faecium strain 64/3xUW2774 plasmid pLG1 hypothetical protein (pLG1-0143) gene, partial cds
GI5	ATCC43255 GL003456		Streptococcus anginosus subsp. whileyi MAS624 DNA
GI5	ATCC43255 GL003457	relaxase/mobilisation protein	Streptococcus anginosus subsp. whileyi MAS624 DNA
GI5	ATCC43255 GL003458	Superfamily II helicase	Campylobacter hominis ATCC BAA-381
GI5	ATCC43255 GL003459		Streptobacillus moniliformis DSM 12112
GI5	ATCC43255 GL003460	Type I site-specific restriction-modification system, R (restriction) subunit and related helicases	Streptobacillus moniliformis DSM 12112
GI5	ATCC43255 GL003461	Restriction endonuclease S subunits	Streptobacillus moniliformis DSM 12112
GI5	ATCC43255 GL003462	Type I restriction-modification system methyltransferase subunit	Streptobacillus moniliformis DSM 12112
GI5	ATCC43255 GL003463	transcriptional regulator	Streptococcus intermedius C270
GI5	ATCC43255 GL003464		Streptococcus anginosus subsp. whileyi MAS624 DNA
GI5	ATCC43255	DNA methylase	Streptococcus anginosus C238

	GL003465		
GI5	ATCC43255 GL003466	O-Methyltransferase involved in polyketide biosynthesis	Treponema pedis str. T A4
GI5	ATCC43255 GL003467	O-Methyltransferase involved in polyketide biosynthesis	Treponema denticola ATCC 35405
GI5	ATCC43255 GL003468		Streptococcus anginosus C238
GI5	ATCC43255 GL003469	conjugative transposon regulatory protein	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI5	ATCC43255 GL003470	Topoisomerase IA	Streptococcus anginosus C238
GI5	ATCC43255 GL003471	Topoisomerase IA	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI5	ATCC43255 GL003472	cell surface protein	Streptococcus dysgalactiae subsp. equisimilis ATCC 12394
GI5	ATCC43255 GL003473		
GI5	ATCC43255 GL003474	Cell wall-associated hydrolases (invasion-associated proteins)	Streptococcus intermedius B196
GI5	ATCC43255 GL003475	Type IV secretory pathway, VirB4 components	Streptococcus anginosus C238
GI5	ATCC43255 GL003476		Schistosoma mansoni hypothetical protein (Smp_090990) mRNA, complete cds
GI5	ATCC43255 GL003477		Streptococcus anginosus subsp. whileyi MAS624 DNA
GI5	ATCC43255 GL003478	single-strand binding protein	Streptococcus equi subsp. equi 4047
GI5	ATCC43255 GL003479	single-stranded DNA binding protein	Streptococcus intermedius C270
GI5	ATCC43255 GL003480	AraC family transcription regulator	Streptococcus parasanguinis FW213
GI5	ATCC43255	Type IV secretory	Streptococcus anginosus C238

	GL003481	pathway, VirD4 components	
GI5	ATCC43255 GL003482		<i>Streptococcus equi</i> subsp. <i>equi</i> 4047
GI5	ATCC43255 GL003483	conjugative transposon protein	<i>Filifactor alocis</i> ATCC 35896
GI5	ATCC43255 GL003484		<i>Ethanoligenens harbinense</i> YUAN-3
GI5	ATCC43255 GL003485	membrane protein	<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i> ATCC 35246
GI5	ATCC43255 GL003486	SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase	<i>Haemophilus influenzae</i> R2846
GI5	ATCC43255 GL003487	Site-specific recombinases, DNA invertase Pin homologs	<i>Faecalibacterium prausnitzii</i> SL3/3 draft genome
GI5	ATCC43255 GL003488		
GI5	ATCC43255 GL003489	conjugative transposon protein	<i>Ruminococcus obeum</i> A2-162 draft genome
GI5	ATCC43255 GL003490		
GI5	ATCC43255 GL003491	Cation transport ATPase	<i>Clostridium phytofermentans</i> ISDg
GI5	ATCC43255 GL003492	mgtC; magnesium transporting ATPase protein C	<i>Eubacterium limosum</i> KIST612
GI5	ATCC43255 GL003493	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	<i>Eubacterium limosum</i> KIST612
GI5	ATCC43255 GL003494	Cation transport ATPase	<i>Eubacterium limosum</i> KIST612

GI5	ATCC43255 GL003495		Clostridiales sp. SSC/2 draft genome
GI5	ATCC43255 GL003496		Clostridiales sp. SSC/2 draft genome
GI5	ATCC43255 GL003497	Cell wall-associated hydrolases (invasion-associated proteins)	Roseburia intestinalis M50/1 draft genome
GI5	ATCC43255 GL003498		
GI5	ATCC43255 GL003499		Uncultured bacterium clone LM0ACA12ZE03FM1 genomic sequence
GI5	ATCC43255 GL003500	SAM-dependent methyltransferases related to tRNA (uracil-5-)-methyltransferase	Alkaliphilus metalliredigens QYMF
GI6	CF5GL0003 88	membrane protein	Streptococcus equi subsp. zooepidemicus ATCC 35246
GI6	CF5GL0003 89		Streptococcus agalactiae A909
GI6	CF5GL0003 90	conjugative transposon protein	Filifactor alocis ATCC 35896
GI6	CF5GL0003 91	conjugative transposon protein	Streptococcus anginosus C238
GI6	CF5GL0003 92	Type IV secretory pathway, VirD4 components	Streptococcus intermedius B196
GI6	CF5GL0003 93	AraC-type DNA-binding domain-containing proteins	Streptococcus anginosus subsp. whileyi MAS624 DNA
GI6	CF5GL0003 94	AraC family transcription regulator	Streptococcus intermedius B196
GI6	CF5GL0003 95	single-stranded DNA binding protein	Streptococcus anginosus C238
GI6	CF5GL0003 96	single-strand binding protein	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA

GI6	CF5GL0003 97	conjugative transposon membrane protein	<i>Streptococcus anginosus</i> subsp. <i>whileyi</i> MAS624 DNA
GI6	CF5GL0003 98	conjugative transposon membrane exported protein	<i>Schistosoma mansoni</i> hypothetical protein (Smp_090990) mRNA, complete cds
GI6	CF5GL0003 99	Type IV secretory pathway, VirB4 components	<i>Streptococcus anginosus</i> C238
GI6	CF5GL0004 00	Cell wall-associated hydrolases (invasion-associ ated proteins)	<i>Streptococcus intermedius</i> B196
GI6	CF5GL0004 01		
GI6	CF5GL0004 02	cell surface protein	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> ATCC 12394
GI6	CF5GL0004 03	Topoisomerase IA	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> RE378 DNA
GI6	CF5GL0004 04	conjugative transposon regulatory protein	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> RE378 DNA
GI6	CF5GL0004 05		<i>Streptococcus anginosus</i> C238
GI6	CF5GL0004 06	O-Methyltransfer ase involved in polyketide biosynthesis	<i>Treponema denticola</i> ATCC 35405
GI6	CF5GL0004 07	DNA methylase	<i>Streptococcus anginosus</i> C238
GI6	CF5GL0004 08		<i>Streptococcus anginosus</i> C238
GI6	CF5GL0004 09	transcriptional regulator	<i>Campylobacter hominis</i> ATCC BAA-381
GI6	CF5GL0004 10		<i>Fusobacterium nucleatum</i> subsp. <i>nucleatum</i> ATCC 25586
GI6	CF5GL0004 11	cytoplasmic protein	<i>Fusobacterium nucleatum</i> subsp. <i>nucleatum</i> ATCC 25586
GI6	CF5GL0004 12	conjugative transposon mobilization	<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C1050

		protein	
GI6	CF5GL0004 13		Clostridiales sp. SSC/2 draft genome
GI6	CF5GL0004 14	Polyferredoxin	Clostridiales sp. SSC/2 draft genome
GI6	CF5GL0004 15	ABC-type antimicrobial peptide transport system, permease component	Clostridiales sp. SSC/2 draft genome
GI6	CF5GL0004 16	ABC-type antimicrobial peptide transport system, ATPase component	Clostridiales sp. SSC/2 draft genome
GI6	CF5GL0004 17	ABC-type antimicrobial peptide transport system, permease component	Enterococcus faecalis 62
GI6	CF5GL0004 18	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Enterococcus faecium Aus0085 plasmid p1, complete sequence
GI6	CF5GL0004 19	Signal transduction histidine kinase	Enterococcus faecalis 62
GI6	CF5GL0004 20		Streptococcus equi subsp. zooepidemicus H70
GI6	CF5GL0004 21	sigma factor	Streptococcus intermedius C270
GI6	CF5GL0004 22		Streptococcus constellatus subsp. pharyngis C1050
GI6	CF5GL0004 23	Site-specific recombinases, DNA invertase Pin	Streptococcus dysgalactiae subsp. equisimilis AC-2713

		homologs	
GI6	CF5GL0004 24	Site-specific recombinases, DNA invertase Pin homologs	Dehalobacter sp. CF
GI6	CF5GL0004 25		
GI6	CF5GL0004 26	Acetyltransferas es, including N-acetylases of ribosomal proteins	Citrobacter rodentium ICC168
GI6	CF5GL0004 27		Sebaldella termitidis ATCC 33386
GI6	CF5GL0004 28	Crp/Fnr family transcriptional regulator	Sebaldella termitidis ATCC 33386
GI7	M68GL0003 79		Streptococcus equi subsp. zooepidemicus ATCC 35246
GI7	M68GL0003 80		Streptococcus agalactiae A909
GI7	M68GL0003 81	conjugative transposon protein	Filifactor alocis ATCC 35896
GI7	M68GL0003 82	conjugative transposon protein	Streptococcus anginosus C238
GI7	M68GL0003 83	Type IV secretory pathway, VirD4 components	Streptococcus intermedius B196
GI7	M68GL0003 84	AraC-type DNA-binding domain-contains g proteins	Streptococcus anginosus subsp. whileyi MAS624 DNA
GI7	M68GL0003 85	AraC family transcription regulator	Streptococcus intermedius B196
GI7	M68GL0003 86	single-stranded DNA binding protein	Streptococcus anginosus C238
GI7	M68GL0003 87	conjugative transposon membrane protein	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI7	M68GL0003	conjugative	Streptococcus anginosus subsp. whileyi

	88	transposon membrane protein	MAS624 DNA
GI7	M68GL0003 89		Schistosoma mansoni hypothetical protein (Smp_090990) mRNA, complete cds
GI7	M68GL0003 90	Type IV secretory pathway, VirB4 components	Streptococcus anginosus C238
GI7	M68GL0003 91	Cell wall-associated hydrolases (invasion-associ ated proteins)	Streptococcus intermedius B196
GI7	M68GL0003 92		
GI7	M68GL0003 93	cell surface protein	Streptococcus dysgalactiae subsp. equisimilis ATCC 12394
GI7	M68GL0003 94	Topoisomerase IA	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI7	M68GL0003 95	conjugative transposon regulatory protein	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI7	M68GL0003 96		Streptococcus anginosus C238
GI7	M68GL0003 97	O-Methyltransfer ase involved in polyketide biosynthesis	Treponema denticola ATCC 35405
GI7	M68GL0003 98	DNA methylase	Streptococcus anginosus C238
GI7	M68GL0003 99		Streptococcus anginosus C238
GI7	M68GL0004 00	transcriptional regulator	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI7	M68GL0004 01		Fusobacterium nucleatum subsp. nucleatum ATCC 25586
GI7	M68GL0004 02	cytoplasmic protein	Fusobacterium nucleatum subsp. nucleatum ATCC 25586
GI7	M68GL0004 03	conjugative transposon mobilization protein	Streptococcus constellatus subsp. pharyngis C1050
GI7	M68GL0004 04		Clostridiales sp. SSC/2 draft genome

GI7	M68GL0004 05	Polyferredoxin	Clostridiales sp. SSC/2 draft genome
GI7	M68GL0004 06	ABC-type antimicrobial peptide transport system, permease component	Clostridiales sp. SSC/2 draft genome
GI7	M68GL0004 07	ABC-type antimicrobial peptide transport system, ATPase component	Clostridiales sp. SSC/2 draft genome
GI7	M68GL0004 08	ABC-type antimicrobial peptide transport system, permease component	Enterococcus faecalis 62
GI7	M68GL0004 09	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Enterococcus faecium Aus0085 plasmid p1, complete sequence
GI7	M68GL0004 10	Signal transduction histidine kinase	Enterococcus faecalis 62
GI7	M68GL0004 11		Streptococcus equi subsp. zooepidemicus H70
GI7	M68GL0004 12	sigma factor	Streptococcus intermedius C270
GI7	M68GL0004 13		Streptococcus constellatus subsp. pharyngis C1050
GI7	M68GL0004 14	Site-specific recombinases, DNA invertase Pin homologs	Streptococcus dysgalactiae subsp. equisimilis AC-2713
GI7	M68GL0004 15	Site-specific recombinases,	Dehalobacter sp. CF

		DNA invertase Pin homologs	
GI7	M68GL0004 16	Site-specific recombinases, DNA invertase Pin homologs	Dehalobacter sp. CF
GI7	M68GL0004 17		
GI7	M68GL0004 18	Acetyltransferases, including N-acetylases of ribosomal proteins	Citrobacter rodentium ICC168
GI7	M68GL0004 19		Sebaldella termitidis ATCC 33386
GI7	M68GL0004 20	Crp/Fnr family transcriptional regulator	Sebaldella termitidis ATCC 33386
GI7	M68GL0004 21	transposase and inactivated derivatives	Enterococcus faecalis plasmid pTW9 DNA, complete sequence
GI7	M68GL0004 22		Ruminococcus bromii L2-63 draft genome
GI8	M120GL000 359	transcriptional regulator	Corynebacterium diphtheriae HC02
GI8	M120GL000 360	DNA modification methylase	Bacillus cereus ATCC 10987
GI8	M120GL000 361	DNA modification methylase	Bacillus cereus ATCC 10987
GI8	M120GL000 362		Bacillus cellulosilyticus DSM 2522
GI8	M120GL000 363	GTPase subunit of restriction endonuclease	Gardnerella vaginalis 409-05
GI8	M120GL000 364	LlaJI restriction endonuclease	Gardnerella vaginalis 409-05
GI8	M120GL000 365		Ruminococcus albus 7
GI8	M120GL000 366	ECF subfamily RNA polymerase sigma-24 factor	Mahella australiensis 50-1 BON
GI8	M120GL000 367		Thermoanaerobacter sp. X513

GI8	M120GL000 368	rRNA biogenesis protein rrp5	Thermoanaerobacter sp. X513
GI8	M120GL000 369		Thermoanaerobacter sp. X513
GI8	M120GL000 370		Thermoanaerobacter sp. X513
GI8	M120GL000 371	DNA-directed DNA polymerase	Thermoanaerobacter sp. X513
GI8	M120GL000 372	Prophage antirepressor	Thermoanaerobacter sp. X513
GI8	M120GL000 373		Thermoanaerobacter sp. X513
GI8	M120GL000 374	P-loop ATPase and inactivated derivatives	Thermoanaerobacter sp. X513
GI8	M120GL000 375	nuclease p44	Thermoanaerobacter sp. X513
GI8	M120GL000 376	Superfamily II DNA/RNA helicases	Thermoanaerobacter sp. X513
GI8	M120GL000 377	phage-associated protein	Thermoanaerobacter sp. X513
GI8	M120GL000 378		Thermoanaerobacter sp. X513
GI8	M120GL000 379	S-adenosylmethio nine synthetase	Thermoanaerobacter sp. X513
GI8	M120GL000 380	DNA modification methylase	Thermoanaerobacter sp. X513
GI8	M120GL000 381	virulence-relate d protein	Thermoanaerobacter sp. X513
GI8	M120GL000 382		Thermoanaerobacter sp. X513
GI8	M120GL000 383	AIG2 family protein	Thermoanaerobacter sp. X513
GI8	M120GL000 384		Thermoanaerobacter sp. X513
GI8	M120GL000 385	Phage terminase-like protein, large subunit	Thermoanaerobacter sp. X513
GI8	M120GL000 386		Streptococcus constellatus subsp. pharyngis C1050
GI8	M120GL000 387	Phage-related protein	Thermoanaerobacter sp. X513

GI8	M120GL000388	Protease subunit of ATP-dependent Clp proteases	Thermoanaerobacter sp. X513
GI8	M120GL000389	phage phi-C31 gp36 major capsid-like protein	Thermoanaerobacter sp. X513
GI8	M120GL000390		Thermoanaerobacter sp. X513
GI8	M120GL000391	phage head-tail adaptor, putative	Thermoanaerobacter sp. X513
GI8	M120GL000392	HK97 family phage protein	Thermoanaerobacter sp. X513
GI8	M120GL000393	phil3 family phage major tail protein	Thermoanaerobacter sp. X513
GI8	M120GL000394	Phage-related protein	Thermoanaerobacter sp. X513
GI8	M120GL000395	Phage-related protein	Thermoanaerobacter sp. X513
GI8	M120GL000396	Phage-related protein	Thermoanaerobacter sp. X513
GI8	M120GL000397		Thermoanaerobacter sp. X513
GI8	M120GL000398		Thermoanaerobacter sp. X513
GI8	M120GL000399	glycosyl hydrolase	Thermoanaerobacter sp. X513
GI8	M120GL000400	Phage-related holin (Lysis protein)	Thermoanaerobacter sp. X513
GI8	M120GL000401	N-acetylmuramoyl-L-alanine amidase	Thermoanaerobacter sp. X513
GI8	M120GL000402		Thermoanaerobacter sp. X513
GI8	M120GL000403	Site-specific recombinases, DNA invertase Pin homologs	Clostridium kluyveri NBRC 12016 DNA
GI8	M120GL000404	phage integrase family site-specific	Streptococcus mitis B6 complete genome, strain B6

		recombinase	
GI8	M120GL000405	Site-specific recombinases, DNA invertase Pin homologs	Streptococcus mitis B6 complete genome, strain B6
GI8	M120GL000406		Staphylococcus aureus Bmb9393
GI8	M120GL000407	nucleotidyltransferase	Staphylococcus aureus Bmb9393
GI8	M120GL000408	SAM-dependent methyltransferases	Enterococcus faecium Aus0085 plasmid p3, complete sequence
GI8	M120GL000409	aadE; streptomycin adenylyltransferase	Staphylococcus aureus strain SA7037 plasmid pV7037, partial sequence
GI8	M120GL000410	Adenine/guanine phosphoribosyltransferases and related PRPP-binding proteins	Staphylococcus aureus strain SA7037 plasmid pV7037, partial sequence
GI8	M120GL000411	nucleotidyltransferases	Staphylococcus aureus strain SA7037 plasmid pV7037, partial sequence
GI8	M120GL000412	replication initiator protein A (RepA) N-terminal domain protein	Streptococcus agalactiae ILRI112 complete genome
GI8	M120GL000413	DNA replication protein	Streptococcus pyogenes MGAS10750
GI8	M120GL000414		Streptococcus agalactiae ILRI112 complete genome
GI8	M120GL000415	prophage antirepressor	Streptococcus pneumoniae AP200
GI8	M120GL000416	Type IV secretory pathway, VirD4 components	Anaerococcus prevotii DSM 20548 plasmid pAPRE01, complete sequence
GI8	M120GL000417	transcriptional regulators	Clostridium saccharolyticum WM1
GI8	M120GL000418		
GI8	M120GL000419	permeases	complete chromosome Acholeplasma brassicae

GI8	M120GL000420	Thiol-disulfide isomerase and thioredoxins	<i>Clostridium clariflavum</i> DSM 19732
GI8	M120GL000421	Lactoylglutathione lyase and related lyases	<i>Campylobacter fetus</i> subsp. <i>fetus</i> genomic DNA containing type IV secretion system and antibiotic resistance gene cluster, strain IMD 523
GI8	M120GL000422	transcriptional regulators	<i>Campylobacter fetus</i> subsp. <i>fetus</i> genomic DNA containing type IV secretion system and antibiotic resistance gene cluster, strain IMD 523
GI8	M120GL000423	ribosomal tetracycline resistance protein tet	<i>Campylobacter fetus</i> subsp. <i>fetus</i> genomic DNA containing type IV secretion system and antibiotic resistance gene cluster, strain IMD 523
GI8	M120GL000424	aadE; streptomycin aminoglycoside 6-adenyltransferase	<i>Campylobacter fetus</i> subsp. <i>fetus</i> genomic DNA containing type IV secretion system and antibiotic resistance gene cluster, strain IMD 523
GI8	M120GL000425		<i>Anaerococcus prevotii</i> DSM 20548 plasmid pAPRE01, complete sequence
GI8	M120GL000426	replication initiator protein A	<i>Streptococcus agalactiae</i> ILRI112 complete genome
GI8	M120GL000427	DNA replication protein	<i>Streptococcus agalactiae</i> ILRI112 complete genome
GI8	M120GL000428		<i>Anaerococcus prevotii</i> DSM 20548 plasmid pAPRE01, complete sequence
GI8	M120GL000429	TnpX site-specific recombinase	<i>Filifactor alocis</i> ATCC 35896
GI8	M120GL000430	Flavodoxins	<i>Filifactor alocis</i> ATCC 35896
GI8	M120GL000431		
GI8	M120GL000432		<i>Streptococcus agalactiae</i> ILRI112 complete genome
GI8	M120GL000433	prophage antirepressor	<i>Streptococcus pneumoniae</i> AP200
GI8	M120GL000434	Type IV secretory pathway, VirD4 components	<i>Anaerococcus prevotii</i> DSM 20548 plasmid pAPRE01, complete sequence
GI8	M120GL000		

	435		
GI8	M120GL000 436		<i>Streptococcus pyogenes</i> MGAS10750
GI8	M120GL000 437	single-strand binding protein	<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> RE378 DNA
GI8	M120GL000 438	conjugative transposon membrane protein	<i>Streptococcus agalactiae</i> ILRI112 complete genome
GI8	M120GL000 439		<i>Finnegoldia magna</i> ATCC 29328 DNA
GI8	M120GL000 440		<i>Streptococcus pyogenes</i> MGAS10750
GI8	M120GL000 441	Type IV secretory pathway, VirB4 components	<i>Finnegoldia magna</i> ATCC 29328 DNA
GI8	M120GL000 442	Cell wall-associated hydrolases (invasion-associated proteins)	<i>Streptococcus pyogenes</i> integrative conjugative element ICESpl108, strain C1
GI8	M120GL000 443		
GI8	M120GL000 444	chimeric erythrocyte-binding protein	<i>Streptococcus pyogenes</i> MGAS10750
GI8	M120GL000 445	bacteriocin	<i>Streptococcus pyogenes</i> MGAS10750
GI8	M120GL000 446		<i>Anaerococcus prevotii</i> DSM 20548 plasmid pAPRE01, complete sequence
GI8	M120GL000 447		<i>Streptococcus agalactiae</i> ILRI112 complete genome
GI8	M120GL000 448	Topoisomerase IA	<i>Streptococcus pyogenes</i> integrative conjugative element ICESpl108, strain C1
GI8	M120GL000 449	transcriptional regulators	Clostridiales genomosp. BVAB3 str. UPII9-5
GI8	M120GL000 450	transporter	Clostridiales genomosp. BVAB3 str. UPII9-5
GI8	M120GL000 451	GNAT domain-containing toxin-antitoxin system toxin protein	Clostridiales genomosp. BVAB3 str. UPII9-5
GI8	M120GL000	Topoisomerase IA	<i>Streptococcus pneumoniae</i> AP200

	452		
GI8	M120GL000 453	Site-specific DNA methylase	<i>Streptococcus pneumoniae</i> AP200
GI8	M120GL000 454	DNA methylase	<i>Streptococcus pyogenes</i> MGAS10750
GI8	M120GL000 455		<i>Aerococcus urinae</i> ACS-120-V-Col110a
GI8	M120GL000 456	transcriptional regulator, XRE family	<i>Anaerococcus prevotii</i> DSM 20548
GI8	M120GL000 457	Permeases of the major facilitator superfamily	<i>Petrotoga mobilis</i> SJ95
GI8	M120GL000 458	relaxase/mobiliz ation nuclease domain protein	<i>Finegoldia magna</i> ATCC 29328 DNA
GI8	M120GL000 459		<i>Streptococcus agalactiae</i> ILRI112 complete genome
GI8	M120GL000 460	Virulence protein	<i>Streptococcus pyogenes</i> integrative conjugative element ICESp1108, strain C1
GI8	M120GL000 461		<i>Streptococcus pyogenes</i> integrative conjugative element ICESp1108, strain C1
GI8	M120GL000 462	Zn peptidase	<i>Streptococcus pneumoniae</i> AP200
GI8	M120GL000 463		<i>Streptococcus pneumoniae</i> AP200
GI8	M120GL000 464		<i>Streptococcus pyogenes</i> MGAS10750
GI8	M120GL000 465	sigma-70, region 4	<i>Streptococcus pneumoniae</i> AP200
GI8	M120GL000 466		<i>Streptococcus agalactiae</i> ILRI112 complete genome
GI8	M120GL000 467	Site-specific recombinases, DNA invertase Pin homologs	<i>Streptococcus pyogenes</i> integrative conjugative element ICESp1108, strain C1
GI8	M120GL000 468	cell surface protein	
GI8	M120GL000 469	HNH endonuclease	<i>Bacillus thuringiensis</i> MC28 plasmid pMC429, complete sequence
GI8	M120GL000 470	Transposase and inactivated derivatives	<i>Clostridium clariflavum</i> DSM 19732

GI8	M120GL000471	Transposase and inactivated derivatives	<i>Clostridium clariflavum</i> DSM 19732
GI8	M120GL000472	transposase, is4	<i>Clostridium clariflavum</i> DSM 19732
GI8	M120GL000473		
GI8	M120GL000474	accessory gene regulator	<i>Clostridium acidurici</i> 9a
GI8	M120GL000475	signal transduction protein with a C-terminal ATPase domain	<i>Clostridium acidurici</i> 9a
GI8	M120GL000476	Cation transport ATPase	<i>Eubacterium limosum</i> KIST612
GI8	M120GL000477	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	<i>Clostridium phytofermentans</i> ISDg
GI8	M120GL000478	Cation transport ATPase	<i>Clostridium phytofermentans</i> ISDg
GI8	M120GL000479	Multimeric flavodoxin WrbA	<i>Halobacteroides halobius</i> DSM 5150
GI8	M120GL000480		
GI8	M120GL000481	L-rhamnose mutarotase	<i>Clostridium phytofermentans</i> ISDg
GI8	M120GL000482	AraC family transcriptional regulator	<i>Paenibacillus polymyxa</i> M1 main chromosome
GI8	M120GL000483	Response regulator containing CheY-like receiver domain and AraC-type DNA-binding domain	<i>Paenibacillus mucilaginosus</i> KNP414

GI8	M120GL000 484	AraC-type DNA-binding domain-contains g proteins	<i>Clostridium saccharoperbutylacetonicum</i> N1-4 (HMT)
GI8	M120GL000 485	Cystathionine beta-lyases/cyst athionine gamma-synthases	<i>Brachyspira pilosicoli</i> WesB complete genome
GI8	M120GL000 486		<i>Clostridium beijerinckii</i> NCIMB 8052
GI9	BJ08GL000 303	membrane protein	<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i> ATCC 35246
GI9	BJ08GL000 304		<i>Streptococcus agalactiae</i> A909
GI9	BJ08GL000 305	conjugative transposon protein	<i>Filifactor alocis</i> ATCC 35896
GI9	BJ08GL000 306	conjugative transposon protein	<i>Streptococcus anginosus</i> C238
GI9	BJ08GL000 307	Site-specific recombinases, DNA invertase Pin homologs	Uncultured organism clone 22 genomic sequence
GI9	BJ08GL000 308		<i>Ruminococcus torques</i> L2-14 draft genome
GI9	BJ08GL000 309		<i>Slackia heliotrinireducens</i> DSM 20476
GI9	BJ08GL000 310		
GI9	BJ08GL000 311	ATP-dependent exoDNase (exonuclease V), alpha subunit - helicase superfamily I member	<i>Ruminococcus torques</i> L2-14 draft genome
GI9	BJ08GL000 312		<i>Clostridiales</i> sp. SS3/4 draft genome
GI9	BJ08GL000 313		<i>Clostridium saccharolyticum</i> -like K10 draft genome
GI9	BJ08GL000 314	DNA primase (bacterial type)	<i>Eubacterium rectale</i> DSM 17629 draft genome
GI9	BJ08GL000	Type IV secretory	<i>Streptococcus pyogenes</i> ICESp2905 DNA

	315	pathway, VirD4 components	containing erm(TR)-carrying element and tet(0) fragment, strain iB21
GI9	BJ08GL000 316	AraC-type DNA-binding domain-containing proteins	<i>Streptococcus anginosus</i> subsp. <i>whileyi</i> MAS624 DNA
GI9	BJ08GL000 317	AraC family transcription regulator	<i>Streptococcus intermedius</i> B196
GI9	BJ08GL000 318	single-stranded DNA binding protein	<i>Streptococcus anginosus</i> C238
GI9	BJ08GL000 319	conjugative transposon membrane protein	<i>Streptococcus constellatus</i> subsp. <i>pharyngis</i> C818
GI9	BJ08GL000 320	transcriptional regulators	<i>Streptococcus gallolyticus</i> subsp. <i>gallolyticus</i> ATCC 43143 DNA
GI9	BJ08GL000 321	sigma-70	<i>Clostridium saccharolyticum</i> -like K10 draft genome
GI9	BJ08GL000 322	Dimethyladenosine transferase (rRNA methylation)	<i>Clostridium acidurici</i> 9a
GI9	BJ08GL000 323	Peptidase E	<i>Faecalibacterium prausnitzii</i> L2/6 draft genome
GI9	BJ08GL000 324		<i>Clostridium saccharolyticum</i> -like K10 draft genome
GI9	BJ08GL000 325		<i>Clostridium</i> sp. SY8519 DNA
GI9	BJ08GL000 326		<i>Clostridium</i> sp. SY8519 DNA
GI9	BJ08GL000 327		<i>Clostridium</i> sp. SY8519 DNA
GI9	BJ08GL000 328	replication-associated protein RepA	<i>Clostridium</i> sp. SY8519 DNA
GI9	BJ08GL000 329	replicative DNA helicase	<i>Clostridium</i> sp. SY8519 DNA
GI9	BJ08GL000 330	Site-specific recombinases, DNA invertase Pin homologs	<i>Treponema succinifaciens</i> DSM 2489
GI9	BJ08GL000 331		<i>Eubacterium rectale</i> M104/1 draft genome

GI9	BJ08GL000332	ribosomal RNA methyltransferase	Citrobacter freundii strain Q1174 class 1 integron OXA-like beta-lactamase (blaOXA-like) gene, partial cds, and aminoglycoside-6' N-acetyltransferase (aacA4), quaternary ammonium compound resistance protein (qacEdelta1), and dihydropteroate synthase type I (sul1) genes, complete cds; insertion sequence ISCR14 putative recombinase ORF494 gene, complete cds; 16S rRNA methyltransferase (rmtD2) and putative tRNA ribosyltransferase genes, complete cds; delta groEL gene, complete sequence; insertion sequence ISCR14b putative recombinase ORF494b gene, complete cds; and hypothetical protein (orf1) gene, partial cds
GI9	BJ08GL000333	methyltransferases	Dictyoglomus turgidum DSM 6724
GI9	BJ08GL000334		Faecalibacterium prausnitzii SL3/3 draft genome
GI9	BJ08GL000335		Faecalibacterium prausnitzii SL3/3 draft genome
GI9	BJ08GL000336	transcriptional regulator, XRE family	Faecalibacterium prausnitzii SL3/3 draft genome
GI9	BJ08GL000337		Faecalibacterium prausnitzii SL3/3 draft genome
GI9	BJ08GL000338		Clostridium beijerinckii NCIMB 8052
GI9	BJ08GL000339		Faecalibacterium prausnitzii SL3/3 draft genome
GI9	BJ08GL000340	ATP-dependent exoDNase (exonuclease V), alpha subunit - helicase superfamily I member	Faecalibacterium prausnitzii SL3/3 draft genome
GI9	BJ08GL000341	DNA primase (bacterial type)	Clostridiales sp. SM4/1 draft genome
GI9	BJ08GL000342	P-loop ATPase and inactivated derivatives	Faecalibacterium prausnitzii SL3/3 draft genome

GI9	BJ08GL000343	Site-specific recombinases, DNA invertase Pin homologs	Faecalibacterium prausnitzii SL3/3 draft genome
GI9	BJ08GL000344	conjugative transposon membrane protein	Streptococcus anginosus subsp. whileyi MAS624 DNA
GI9	BJ08GL000345	conjugative transposon membrane exported protein	Schistosoma mansoni hypothetical protein (Smp_090990) mRNA, complete cds
GI9	BJ08GL000346	Type IV secretory pathway, VirB4 components	Streptococcus anginosus C238
GI9	BJ08GL000347	Cell wall-associated hydrolases (invasion-associated proteins)	Streptococcus intermedius B196
GI9	BJ08GL000348		
GI9	BJ08GL000349	cell surface protein	Streptococcus dysgalactiae subsp. equisimilis ATCC 12394
GI9	BJ08GL000350	Topoisomerase IA	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI9	BJ08GL000351	conjugative transposon regulatory protein	Streptococcus dysgalactiae subsp. equisimilis RE378 DNA
GI9	BJ08GL000352		Streptococcus anginosus C238
GI9	BJ08GL000353	O-Methyltransferase involved in polyketide biosynthesis	Treponema denticola ATCC 35405
GI9	BJ08GL000354	DNA methylase	Streptococcus anginosus C238
GI9	BJ08GL000355		Streptococcus anginosus C238
GI9	BJ08GL000356	transcriptional regulator	Campylobacter hominis ATCC BAA-381
GI9	BJ08GL000357		Fusobacterium nucleatum subsp. nucleatum ATCC 25586
GI9	BJ08GL000	cytoplasmic	Fusobacterium nucleatum subsp.

	358	protein	nucleatum ATCC 25586
GI9	BJ08GL000 359	conjugative transposon mobilization protein	Streptococcus constellatus subsp. pharyngis C1050
GI9	BJ08GL000 360		Clostridiales sp. SSC/2 draft genome
GI9	BJ08GL000 361	Polyferredoxin	Clostridiales sp. SSC/2 draft genome
GI9	BJ08GL000 362	ABC-type antimicrobial peptide transport system, permease component	Clostridiales sp. SSC/2 draft genome
GI9	BJ08GL000 363	ABC-type antimicrobial peptide transport system, ATPase component	Clostridiales sp. SSC/2 draft genome
GI9	BJ08GL000 364	ABC-type antimicrobial peptide transport system, permease component	Enterococcus faecalis 62
GI9	BJ08GL000 365	Response regulators consisting of a CheY-like receiver domain and a winged-helix DNA-binding domain	Enterococcus faecium Aus0085 plasmid p1, complete sequence
GI9	BJ08GL000 366	Signal transduction histidine kinase	Enterococcus faecalis 62
GI9	BJ08GL000 367		Streptococcus equi subsp. zooepidemicus H70
GI9	BJ08GL000 368	sigma factor	Streptococcus intermedius C270
GI9	BJ08GL000		Streptococcus constellatus subsp.

	369		pharyngis C1050
GI9	BJ08GL000 370	Site-specific recombinases, DNA invertase Pin homologs	Streptococcus dysgalactiae subsp. equisimilis AC-2713
GI9	BJ08GL000 371	Site-specific recombinases, DNA invertase Pin homologs	Dehalobacter sp. CF
GI9	BJ08GL000 372	Site-specific recombinases, DNA invertase Pin homologs	Dehalobacter sp. CF
GI9	BJ08GL000 373		
GI9	BJ08GL000 374		
GI9	BJ08GL000 375	Acetyltransferas es, including N-acetylases of ribosomal proteins	Citrobacter rodentium ICC168
GI9	BJ08GL000 376		Sebaldella termitidis ATCC 33386
GI9	BJ08GL000 377	Crp/Fnr family transcriptional regulator	Sebaldella termitidis ATCC 33386
GI9	BJ08GL000 378	transposase and inactivated derivatives	Enterococcus faecalis plasmid pTW9 DNA, complete sequence
GI9	BJ08GL000 379		Ruminococcus bromii L2-63 draft genome

Supplemental Table 5. The function category of genes in 10 T4SS GIs. Red represent the existence of gene, while white represent not. The number is the gene number annotated to be this function by comparing with NT, NR, COG and KEGG database.

Function category	G11	G12	G13	G14	G15	G16	G17	G18	G19	G10
ABC-type	4	3	2	2	2	2	2	0	2	5
regulator	1	1	4	2	5	5	5	9	7	5
endonuclease/excisionase	0	1	1	1	1	0	0	3	0	1
Other genes in conjugative transposon	0	4	0	2	2	2	2	0	2	2
ATPase	0	0	0	1	3	0	0	3	1	2
Na ⁺ -driven multidrug efflux pump/resistance gene	1	0	0	0	0	0	0	1	0	1
cell surface protein	1	0	1	1	1	1	1	1	1	1
Cell wall-associated hydrolases (invasion-associate	0	0	1	0	2	1	1	1	1	0
mobile/mobilization protein	2	1	2	1	1	1	1	1	1	1
methyltransferase/methylase/helicase	1	1	2	2	8	2	2	7	8	4
exported protein	1	0	1	0	0	1	0	0	1	1
single-stranded DNA binding protein	2	0	1	1	2	2	1	1	1	1
recombinase	1	0	0	0	1	1	2	3	4	5
Topoisomerase IA	1	1	1	1	2	1	1	2	1	1
membrane protein	3	2	3	2	1	1	1	0	2	1
Energy production and conversion	0	0	1	1	1	1	1	2	1	1
phage related protein	0	0	0	0	0	0	0	15	0	2
Other genes related with Replication, recombination	2	3	0	1	0	0	1	9	4	4
sigma factor	1	0	1	1	1	1	1	1	2	2
Signal transduction mechanisms	0	1	1	1	1	1	1	1	1	2
toxin/virulence gene	0	0	0	0	0	0	0	3	0	1
VirB11	0	0	1	1	1	1	1	0	1	1
VirB4	1	1	1	1	1	1	1	1	1	1
VirB6	1	1	1	1	1	1	1	1	1	1
VirD2	0	1	0	0	0	0	0	0	0	0
VirD4	1	1	1	2	1	1	1	2	2	0

Supplemental Figure 1. GC content on genome level of 10 *C. difficile* strains

Supplemental Figure 2. The phylogenetic tree based on Topoisomerase IA gene.

Supplemental Figure 3. Gene source in 10 T4SS GIs on genus level.