# The *Clostridioides difficile* species problem: global phylogenomic analysis uncovers three ancient, toxigenic, genomospecies

Daniel R. Knight[1], Korakrit Imwattana[2,3], Brian Kullin[4], Enzo Guerrero-Araya[5,6], Daniel Paredes-Sabja[5,6,7], Xavier Didelot[8], Kate E. Dingle[9], David W. Eyre[10], César Rodríguez[11], and Thomas V. Riley[1,2,12,13*]

[1] Medical, Molecular and Forensic Sciences, Murdoch University, Murdoch, Western Australia, Australia. [2] School of Biomedical Sciences, the University of Western Australia, Nedlands, Western Australia, Australia. [3] Department of Microbiology, Faculty of Medicine Siriraj Hospital, Mahidol University, Thailand. [4] Department of Pathology, University of Cape Town, Cape Town, South Africa. [5] Microbiota-Host Interactions and Clostridia Research Group, Facultad de Ciencias de la Vida, Universidad Andrés Bello, Santiago, Chile. [6] Millenium Nucleus in the Biology of Intestinal Microbiota, Santiago, Chile. [7]Department of Biology, Texas A&M University, College Station, TX, 77843, USA. [8] School of Life Sciences and Department of Statistics, University of Warwick, Coventry, UK. [9] Nuffield Department of Clinical Medicine, University of Oxford, Oxford, UK; National Institute for Health Research (NIHR) Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK. [10] Big Data Institute, Nuffield Department of Population Health, University of Oxford, Oxford, UK; National Institute for Health Research (NIHR) Oxford Biomedical Research Centre, John Radcliffe Hospital, Oxford, UK. [11] Facultad de Microbiología & Centro de Investigación en Enfermedades Tropicales (CIET), Universidad de Costa Rica, San José, Costa Rica. [12] School of Medical and Health Sciences, Edith Cowan University, Joondalup, Western Australia, Australia. [13] Department of Microbiology, PathWest Laboratory Medicine, Queen Elizabeth II Medical Centre, Nedlands, Western Australia, Australia.

*Address correspondence to Professor Thomas V. Riley (thomas.riley@uwa.edu.au), School of Biomedical Sciences, The University of Western Australia, Nedlands, Western Australia, Australia.

Word count (main text): 5090
Abstract word count: 148

## Abstract

*Clostridioides difficile* infection (CDI) remains an urgent global One Health threat. The genetic heterogeneity seen across *C. difficile* underscores its wide ecological versatility and has driven the significant changes in CDI epidemiology seen in the last 20 years. We analysed an international collection of over 12,000 *C. difficile* genomes spanning the eight currently defined phylogenetic clades. Through whole-genome average nucleotide identity, pangenomic and Bayesian analyses, we identified major taxonomic incoherence with clear species boundaries for each of the recently described cryptic clades CI-III. The emergence of these three novel genomospecies predates clades C1-5 by millions of years, rewriting the global population structure of *C. difficile* specifically and taxonomy of the *Peptostreptococcaceae* in general. These genomospecies all show unique and highly divergent toxin gene architecture, advancing our understanding of the evolution of *C. difficile* and close relatives. Beyond the taxonomic ramifications, this work impacts the diagnosis of CDI worldwide.

## Introduction

The bacterial species concept remains controversial, yet it serves as a critical framework for all aspects of modern microbiology[1]. The prevailing species definition describes a genomically coherent group of strains sharing high similarity in many independent phenotypic and ecological properties[2]. The era of whole-genome sequencing (WGS) has seen average nucleotide identity (ANI) replace DNA-DNA hybridization as the 'next-generation' standard for microbial taxonomy[3, 4]. Endorsed by the National Center for Biotechnology Information (NCBI)[4], ANI provides a precise, objective and scalable method for delineation of species, defined as monophyletic groups of strains with genomes that exhibit at least 96% ANI[5, 6].

*Clostridioides (Clostridium) difficile* is an important gastrointestinal pathogen that places a significant growing burden on health care systems in many regions of the world[7]. In both its 2013[8] and 2019[9] reports on antimicrobial resistance (AMR), the US Centers for Disease Control and Prevention rated *C. difficile* infection (CDI) as an urgent health threat, the highest level. Community-associated CDI has become more frequent[7], likely because *C. difficile* has become established in livestock worldwide, resulting in significant environmental contamination[10]. Thus, over the last two decades, CDI has emerged as an important One Health issue[10].

Based on multi-locus sequence type (MLST), there are eight recognised monophyletic groups or 'clades' of *C. difficile*[11]. Strains within these clades show many unique clinical, microbiological and ecological features[11]. Critical to the pathogenesis of CDI is the expression of the large clostridial toxins, TcdA and TcdB and, in some strains, binary toxin (CDT), encoded by two separate chromosomal loci, the PaLoc and CdtLoc, respectively[12]. Clade 1 (C1) contains over 200 toxigenic and non-toxigenic sequence types (STs) including many of the most prevalent strains causing CDI worldwide e.g. ST2, ST8, and ST17[11]. Several highly virulent CDT-producing strains, including ST1 (PCR ribotype (RT) 027), a lineage associated with major hospital outbreaks in North America, Europe and Latin America[13], are found in clade 2 (C2). Comparatively little is known about clade 3 (C3) although it contains ST5 (RT 023), a toxigenic CDT-producing strain with characteristics that may make laboratory detection difficult[14]. *C. difficile* ST37 (RT 017) is found in clade 4 (C4) and, despite the absence of a toxin A gene, is responsible for much of the endemic CDI burden in Asia[15]. Clade 5 (C5) contains several CDT-producing strains including ST11 (RTs 078, 126 and others), which are highly prevalent in production animals worldwide[16]. The remaining so-called 'cryptic' clades (C-I, C-II and C-III), first described in 2012[17, 18], contain over 50 STs from clinical and environmental sources[17, 18, 19, 20, 21]. Evolution of the cryptic clades is poorly understood. Clade C-I strains can cause CDI, however, due to atypical toxin gene architecture, they may not be detected, thus their prevalence may have been underestimated[21].

There are over 600 STs currently described and some STs may have access to a gene pool in excess of 10,000 genes[11, 16, 22]. Considering such enormous diversity, and recent contentious taxonomic revisions[23, 24], we hypothesise that *C. difficile* comprises a complex of distinct species divided along the major evolutionary clades. In this study, whole-genome ANI, and pangenomic and Bayesian analyses are used to explore an international collection of over 12,000 *C. difficile* genomes, to provide new insights into ancestry, genetic diversity and evolution of pathogenicity in this enigmatic pathogen.

# Results

**An updated global population structure based on sequence typing of 12,000 genomes.** We obtained and determined the ST and clade for a collection of 12,621 *C. difficile* genomes (taxid ID 1496, Illumina data) existing in the NCBI Sequence Read Archive (SRA) as of 1st January 2020. A total of 272 STs were identified spanning the eight currently described clades, indicating that the SRA contains genomes for almost 40% of known *C. difficile* STs worldwide (n=659, PubMLST, January 2020). C1 STs dominated the database in both prevalence and diversity (**Fig. 1**) with 149 C1 STs comprising 57.2% of genomes, followed by C2 (35 STs, 22.9%), C5 (18 STs, 10.2%), C4 (34 STs, 7.5%), C3 (7 STs, 2.0%) and the cryptic clades C-I, C-II and C-III (collectively 17 STs, 0.2%). The five most prevalent STs represented were ST1 (20.9% of genomes), ST11 (9.8%), ST2 (9.5%), ST37 (6.5%) and ST8 (5.2%), all prominent lineages associated with CDI worldwide[11].

**Fig. 2** shows an updated global *C. difficile* population structure based on the 659 STs; 27 novel STs were found (an increase of 4%) and some corrections to assignments within C1 and C2 were made, including assigning ST122[25] to C1. Based on PubMLST data and bootstraps values of 1.0 in all monophyletic nodes of the cryptic clades (**Fig. 2**), we could confidently assign 25, 9 and 10 STs to cryptic clades I, II and III, respectively. There remained 26 STs spread across the phylogeny that did not fit within a specific clade (defined as outliers). The tree file for **Fig. 2** and full MLST data is available as **Supplementary Data** at http://doi.org/10.6084/m9.figshare.12471461.

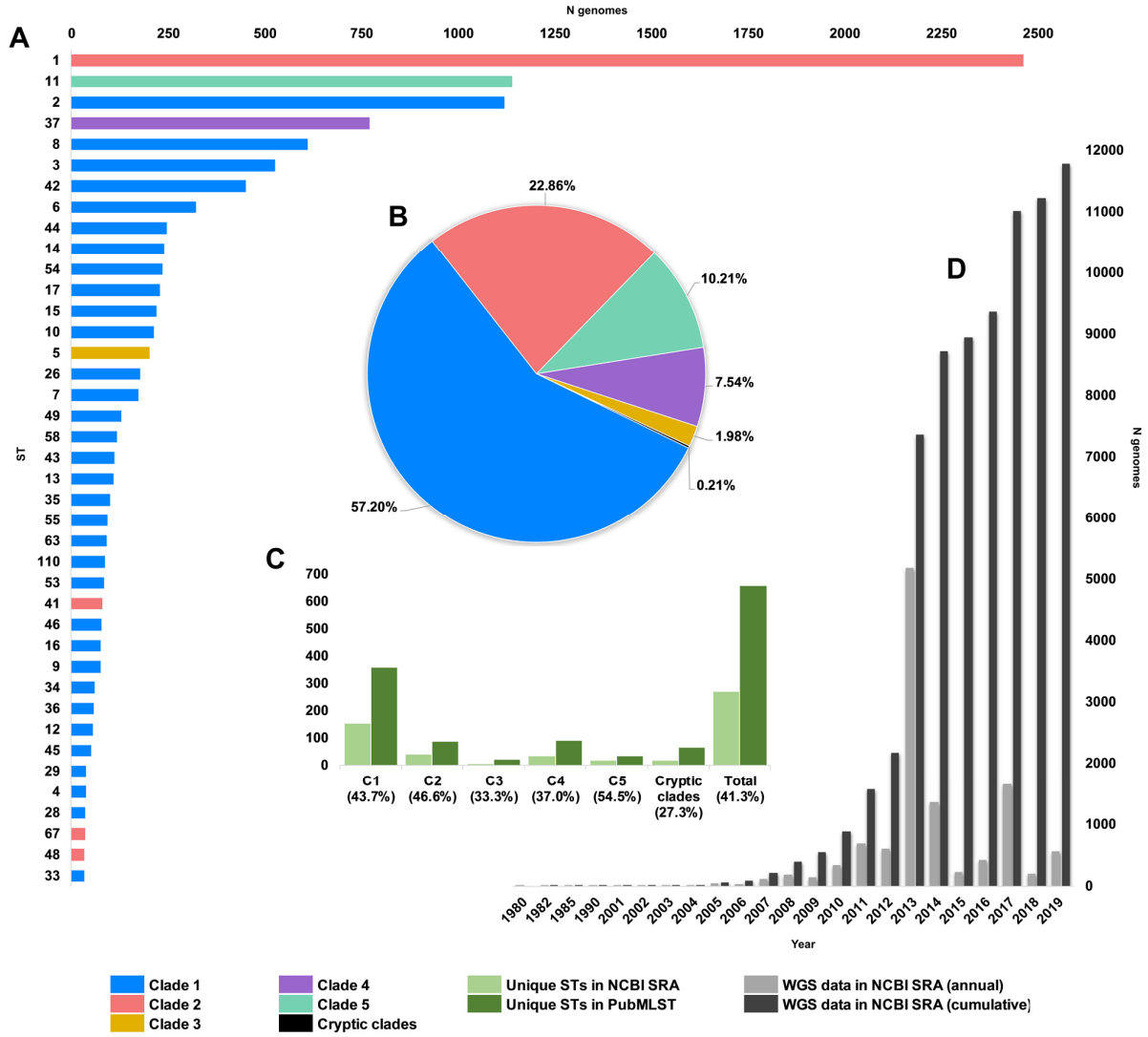


**Figure 1. Composition of *C. difficile* genomes in the NCBI SRA.** Snapshot obtained 1st January 2020; 12,304 strains, [taxid ID 1496]. **(A)** Top 40 most prevalent STs in the NCBI SRA coloured by clade. **(B)** The proportion of genomes in ENA by clade. **(C)** Number/ proportion of STs per clade found in the SRA/present in the PubMLST database. **(D)** Annual and cumulative deposition of *C. difficile* genome data in ENA.

**Whole-genome ANI analysis reveals clear species boundaries.** Whole-genome ANI analyses were used to investigate genetic discontinuity across the *C. difficile* species (**Fig. 3** and **Supplementary Data**). Representative genomes of each ST, chosen based on metadata, read depth and quality, were assembled and annotated. Whole-genome ANI values were determined for a final set of 260 STs using three independent ANI algorithms (FastANI, ANIm and ANIb, see *Methods*). All 225 genomes belonging to clades C1-4 clustered within an ANI range of 97.1-99.8% (median FastANI values of 99.2, 98.7, 97.9 and 97.8%, respectively, **Fig. 3A-C**).

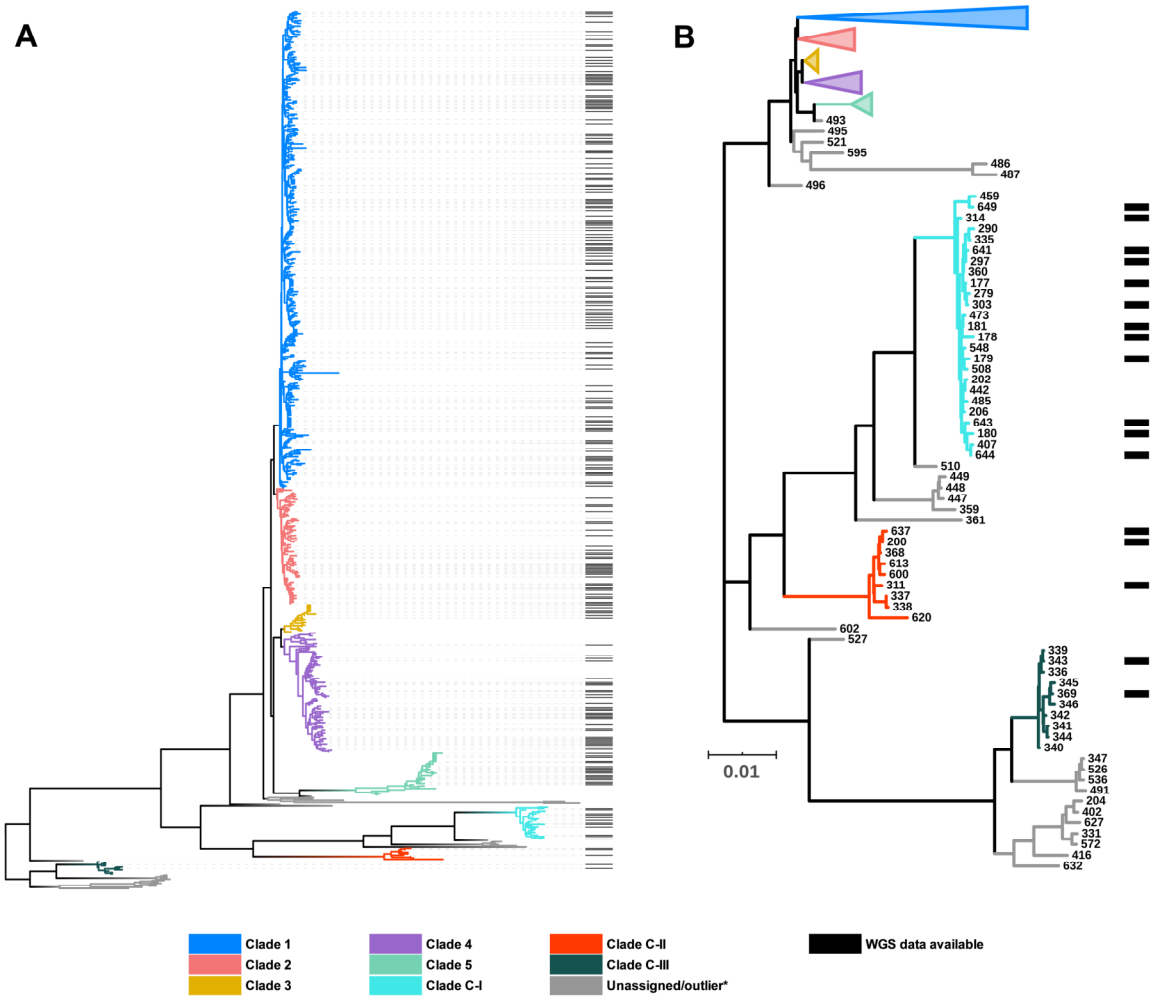


**Figure 2. *C. difficile* population structure. (A)** NJ phylogeny of 659 aligned, concatenated, multilocus sequence type allele combinations coloured by current PubMLST clade assignment. Black bars indicate WGS available for ANI analysis (n=260). **(B)** A subset of the NJ tree showing cryptic clades C-I, C-II and C-III. Again, black bars indicate WGS available for ANI analysis (n=17).

These ANI values are above the 96% species demarcation threshold used by the NCBI[4] and indicate that strains from these clades belong to the same species. ANI values for all 18 genomes belonging to C5 clustered on the borderline of the species demarcation threshold (FastANI range 95.9-96.2%, median 96.1%). ANI values for all three cryptic clades fell well below the species threshold; C-I (FastANI range 90.9-91.1%, median 91.0%), C-II (FastANI range 93.6-93.9%, median 93.7%) and C-III (FastANI range 89.1-89.1%, median 89.1%). All results were corroborated across the three independent ANI algorithms (**Fig. 3A-C**). *C. difficile* strain ATCC 9689 (ST3, C1) was defined by Lawson *et al*. as the type strain for the species[23], and used as a reference in all the above analyses. To better understand the diversity among the divergent clades themselves, FastANI analyses were repeated using STs 11, 181, 200 and 369 as reference archetypes of clades C5, C-I, C-II and C-III, respectively. This approach confirmed that C5 and the three cryptic clades were as distinct from each other as they were collectively from C1-4 (**Fig. 3D-G**).

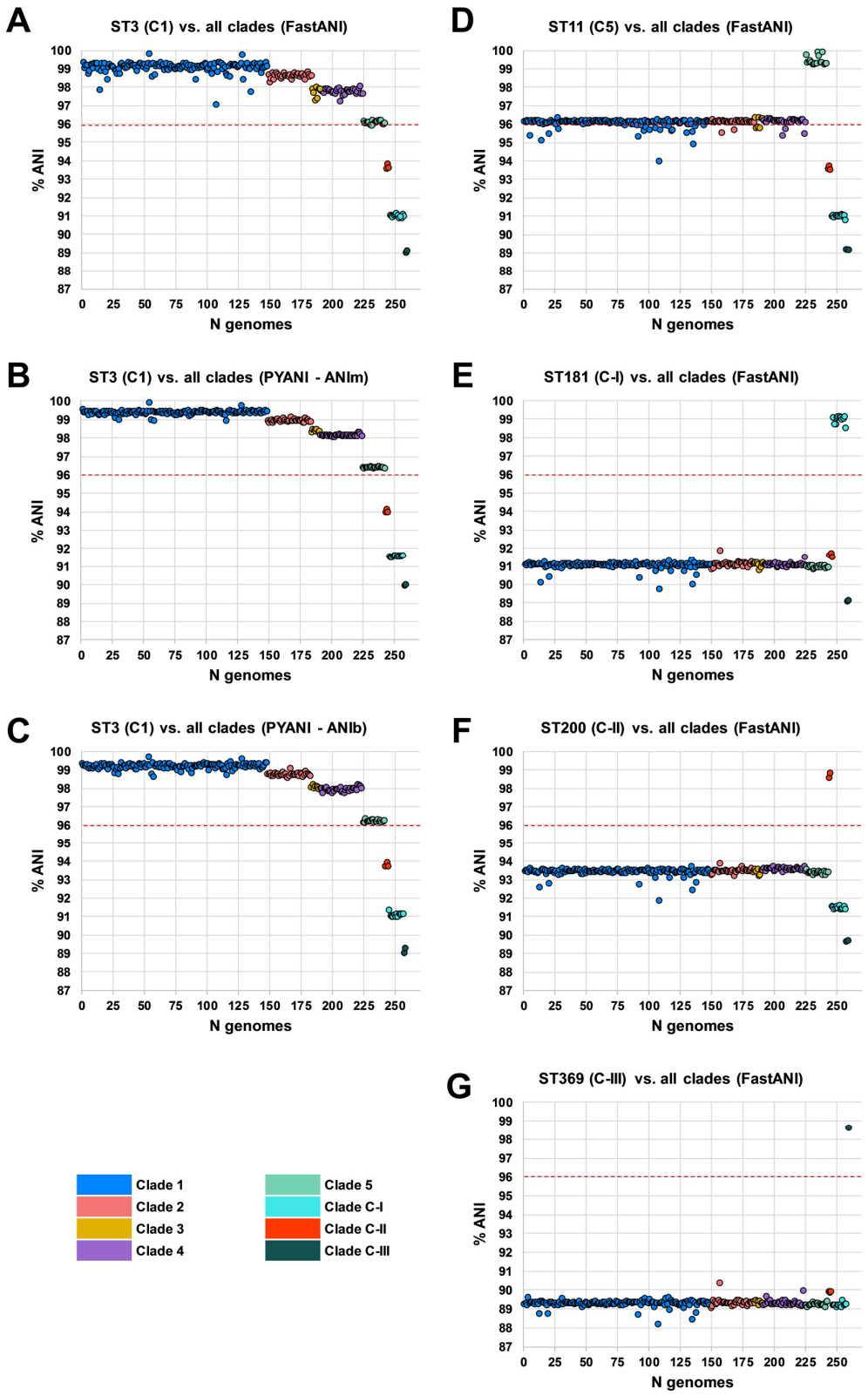


**Figure 3. Species-wide ANI analysis.** Panels **A-C** show ANI plots for ST3 (C1) vs. all clades (260 STs) using FastANI, ANIm and ANIb algorithms, respectively. Panels **D-G** show ANI plots for ST11 (C5), ST181 (C-I), ST200 (C-II) and ST369 (C-III) vs all clades (260 STs), respectively. NCBI species demarcation of 96% indicated by red dashed line[4].

**Taxonomic placement of cryptic clades predates *C. difficile* emergence by millions of years.** Previous studies using BEAST have estimated the common ancestor of C1-5 existed between 1 to 85 or 12 to 14 million years ago (mya)[26, 27]. Here, we used an alternative Bayesian approach, BactDating, to estimate the age of all eight *C. difficile* clades currently described. The last common ancestor for *C. difficile* clades C1-5 was estimated to have existed ~3.89 mya with a 95% credible interval (CI) of 1.11 to 6.71 mya (**Fig. 4**). In contrast, C-II, C-I and C-III emerged 13.05 mya (95% CI 3.72-22.44), 22.02 (95% CI 6.28-37.83) and 47.61 mya (95% CI 13.58-81.73), respectively, at least 9 million years (Megaannum, Ma) before the common ancestor of C1-5. Independent analysis with BEAST, using a smaller core gene dataset (see *Methods*), provided broader estimates of clade emergence, though the emergence order was maintained; C1-5 12.01 mya (95% CI 6.80-33.47), C-II 37.12 mya (95% CI 20.95-103.48), C-I 65.93 mya (95% CI 37.32-183.84) and C-III 142.13 mya (95% CI 79.77-397.18) (**Fig. 4**).

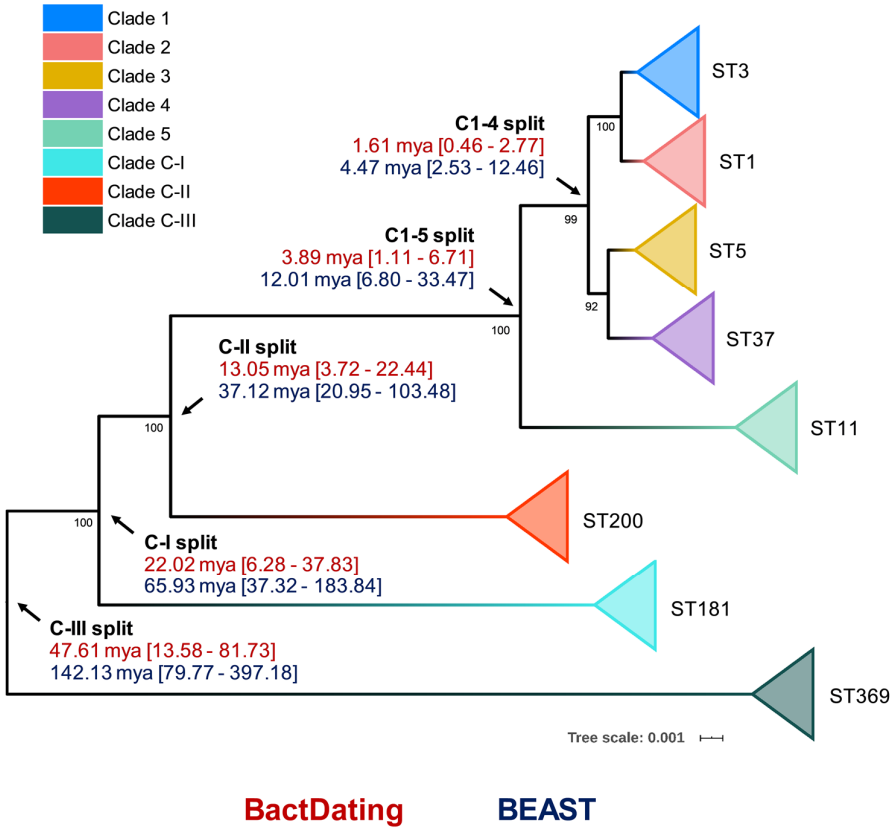


**Figure 4. Bayesian analysis of species and clade divergence.** BactDating and BEAST estimates of the age of major *C. difficile* clades. Node dating ranges for both Bayesian approaches are transposed onto an ML phylogeny built from concatenated MLST alleles of a dozen STs from each clade. Archetypal STs in each evolutionary clade are indicated. The tree is midpoint rooted and bootstrap values are shown. Scale bar indicates the number of substitutions per site. BactDating places the time of most recent common ancestor of C1-5 at 3.89 million years ago (mya) [95% credible interval (CI), 1.11-6.71 mya]. Of the cryptic clades, C-II shared the most recent common ancestor with C1-5 13.05 mya [95% CI 3.72-22.44 mya], followed by C-I (22.02 mya [95% CI 6.28-37.83 mya]), and C-III (47.61 mya [95% CI 13.58-81.73 mya]). Comparative estimates from BEAST are clades C1-5 (12.01 mya [95% CI 6.80-33.47 mya]), C-II (37.12 mya [95% CI 20.95-103.48 mya]), C-I (65.93 mya [95% CI 37.32-183.84 mya]), and C-III (142.13 [95% CI 79.77-397.18 mya]).

153 Next, to identify their true taxonomic placement, ANI was determined for ST181 (C-I), ST200 (C-II)
154 and ST369 (C-III) against two reference datasets. The first dataset comprised 25 species belonging to
155 the *Peptostreptococcaceae* as defined by Lawson *et al*.[23] in their 2016 reclassification of *Clostridium*
156 *difficile* to *Clostridioides difficile*. The second dataset comprised 5,895 complete genomes across 21
157 phyla from the NCBI RefSeq database (accessed 14th January 2020), including 1,366 genomes
158 belonging to *Firmicutes*, 92 genomes belonging to 15 genera within the *Clostridiales* and 20
159 *Clostridium* and *Clostridioides* species. The nearest ANI matches to species within the
160 *Peptostreptococcaceae* dataset were *C. difficile* (range 89.3-93.5% ANI), *Asaccharospora irregularis*
161 (78.9-79.0% ANI) and *Romboutsia lituseburensis* (78.4-78.7% ANI). Notably, *Clostridioides*
162 *mangenotii*, the only other known member of *Clostridioides*, shared only 77.2-77.8% ANI with the
163 cryptic clade genomes (**Table 1**).
164 Similarly, the nearest ANI matches to species within the RefSeq dataset were several
165 *C. difficile* strains (range C-I: 90.9-91.1%; C-II: 93.4-93.6%; and C-III: 89.2-89.4%) and
166 *Paeniclostridium sordellii* (77.7-77.9%). A low ANI (range ≤70-75%) was observed between the
167 cryptic clade genomes and 20 members of the *Clostridium* including *C. tetani, C. botulinum*,
168 *C. perfringens* and *C. butyricum*, the type strain of the *Clostridium* genus *senso stricto*. An updated
169 ANI-based taxonomy for the *Peptostreptococcaceae* is shown in **Fig. 5A**. The phylogeny places C-I,
170 C-II and C-III between *C. mangenotii* and *C. difficile* C1-5, suggesting that they should be assigned
171 to the *Clostridioides* genus, distinct from both *C. mangenotii* and *C. difficile*. Comparative analysis
172 of ANI and 16S rRNA values for the eight *C. difficile* clades and *C. mangenotii* shows significant
173 incongruence between the data generated by the two approaches (**Fig. 5B**). The range of 16S rRNA
174 % similarity between *C. difficile* C1-4, cryptic clades I-III and *C. mangenotii* was narrower (range
175 94.5-100) compared to the range of ANI values (range 77.8-98.7).

176 **Table 1 Whole-genome ANI analysis of cryptic clades vs. 25 *Peptostreptococcaceae* species**
177 **from Lawson *et al*[23].**

| Species | NCBI accession | ANI % | | |
|---|---|---|---|---|
| | | ST181 (C-I) | ST200 (C-II) | ST369 (C-III) |
| *Clostridioides difficile* (ST3) | AQWV00000000.1 | 91.11 | 93.54 | 89.30 |
| *Asaccharospora irregularis* | NZ_FQWX00000000 | 78.94 | 78.87 | 78.91 |
| *Romboutsia lituseburensis* | NZ_FNGW00000000.1 | 78.51 | 78.36 | 78.66 |
| *Romboutsia ilealis* | LN555523.1 | 78.45 | 78.54 | 78.44 |
| *Paraclostridium benzoelyticum* | NZ_LBBT00000000.1 | 77.92 | 77.71 | 78.14 |
| *Paraclostridium bifermentans* | NZ_AVNC00000000.1 | 77.89 | 77.89 | 78.06 |
| *Clostridium mangenotii* | GCA_000687955.1 | 77.82 | 77.84 | 78.15 |
| *Paeniclostridium sordellii* | NZ_APWR00000000.1 | 77.73 | 77.59 | 77.86 |
| *Clostridium hiranonis* | NZ_ABWP01000000 | 77.52 | 77.42 | 77.59 |
| *Terrisporobacter glycolicus* | NZ_AUUB00000000.1 | 77.47 | 77.53 | 77.53 |
| *Intestinibacter bartlettii* | NZ_ABEZ00000000.2 | 77.29 | 77.52 | 77.48 |
| *Clostridium paradoxum* | NZ_LSFY00000000.1 | 76.60 | 76.65 | 76.93 |
| *Clostridium thermoalcaliphilum* | NZ_MZGW00000000.1 | 76.49 | 76.61 | 76.85 |
| *Tepidibacter formicigenes* | NZ_FRAE00000000.1 | 76.41 | 76.47 | 76.38 |
| *Tepidibacter mesophilus* | NZ_BDQY00000000.1 | 76.38 | 76.44 | 76.22 |
| *Tepidibacter thalassicus* | NZ_FQXH00000000.1 | 76.34 | 76.31 | 76.46 |
| *Peptostreptococcus russellii* | NZ_JYGE00000000.1 | 76.30 | 76.08 | 76.38 |
| *Clostridium formicaceticum* | NZ_CP020559.1 | 75.18 | 75.26 | 75.62 |
| *Clostridium caminithermale* | FRAG00000000 | 74.97 | 75.07 | 75.03 |
| *Clostridium aceticum* | NZ_JYHU00000000.1 | ≤70.00 | ≤70.00 | ≤70.00 |
| *Clostridium litorale* | FSRH01000000 | ≤70.00 | ≤70.00 | ≤70.00 |
| *Eubacterium acidaminophilum* | NZ_CP007452.1 | ≤70.00 | ≤70.00 | ≤70.00 |
| *Filifactor alocis* | NC_016630.1 | ≤70.00 | ≤70.00 | ≤70.00 |
| *Peptostreptococcus anaerobius* | ARMA01000000 | ≤70.00 | ≤70.00 | ≤70.00 |
| *Peptostreptococcus stomatis* | NZ_ADGQ00000000.1 | ≤70.00 | ≤70.00 | ≤70.00 |

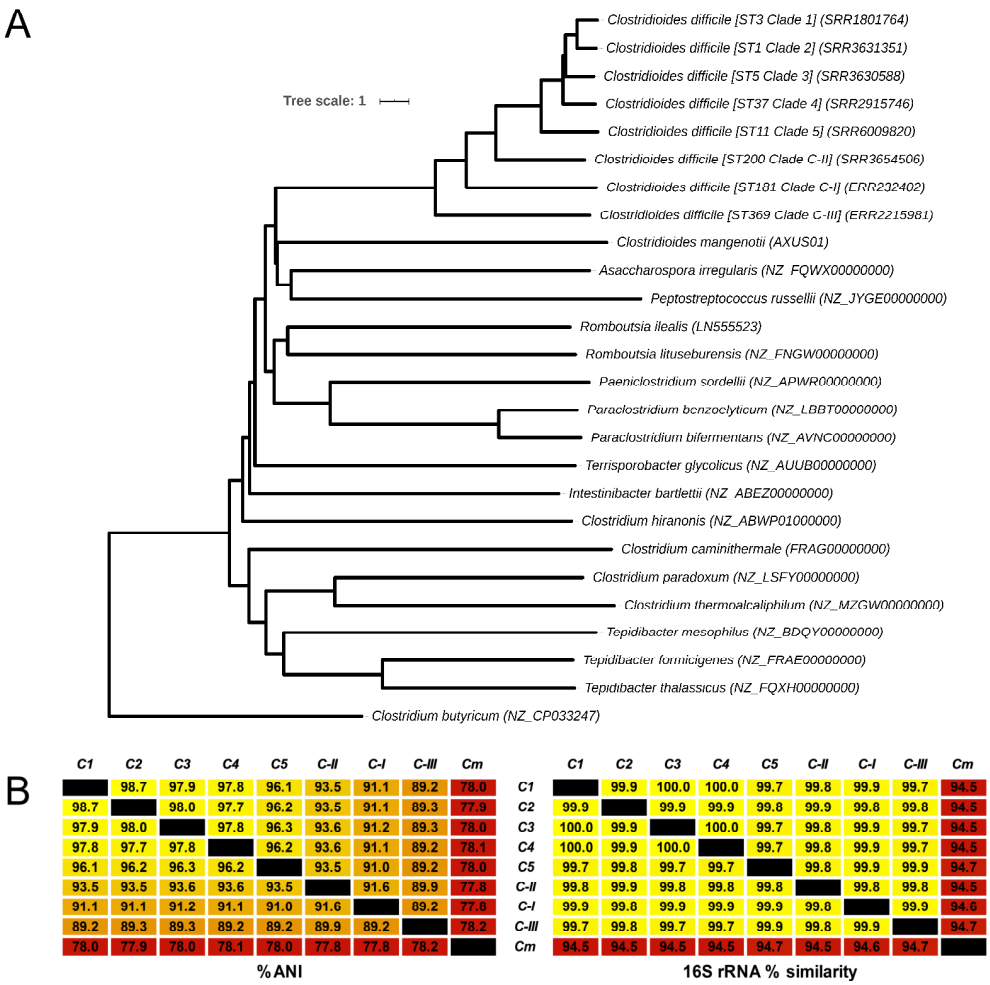


**Figure 5. Revised taxonomy for the *Peptostreptococcaceae*. (A)** ANI-based minimum evolution tree showing evolutionary relationship between eight *C. difficile* 'clades' along with 17 members of the *Peptostreptococcaceae* (from Lawson *et al*[23]) as well as *Clostridium butyricum* as the outgroup and type strain of the *Clostridium* genus *senso stricto*. To convert the ANI into a distance, its complement to 1 was taken. **(B)** Matrices showing pairwise ANI and 16S rRNA values for the eight *C. difficile* clades and *C. mangenotii*, the only other known member of *Clostridioides*.

**Evolutionary and ecological insights from the *C. difficile* species pangenome**. Next, we sought to quantify the *C. difficile* species pangenome and identify genetic loci that are significantly associated with the taxonomically divergent clades. With Panaroo, the *C. difficile* species pangenome comprised 17,470 genes, encompassing an accessory genome of 15,238 genes and a core genome of 2,232 genes, just 12.8% of the total gene repertoire (**Fig 6**). The size of the pangenome reduced by 2,082 genes with the exclusion of clades CI-III, and a further 519 genes with the exclusion of C5. Compared to Panaroo, Roary overestimated the size of the pangenome (32,802 genes), resulting in markedly different estimates of the percentage core genome, 3.9 and 12.8%, respectively ($\chi^2$=1,395.3, df=1, p<0.00001). Panaroo can account for errors introduced during assembly and annotation, thus polishing the 260 Prokka-annotated genomes with Panaroo resulted in a significant reduction in gene content per genome (median 2.48%; 92 genes, range 1.24-12.40%; 82-107 genes, p<0.00001). The *C. difficile* species pangenome was determined to be open[28] (**Fig 6**).

Pan-GWAS analysis with Scoary revealed 142 genes with significant clade specificity. Based on KEGG orthology, these genes were classified into four functional categories: environmental

information processing (7), genetic information processing (39), metabolism (43), and signalling and cellular processes (53). We identified several uniquely present, absent or organised gene clusters associated with ethanolamine catabolism (C-III), heavy metal uptake (C-III), polyamine biosynthesis (C-III), fructosamine utilisation (C-I, C-III), zinc transport (C-II, C5) and folate metabolism (C-I, C5). A summary of the composition and function of these major lineage-specific gene clusters is given in **Table 2,** and a comparative analysis of their respective genetic architecture can be found in the **Supplementary Data**.

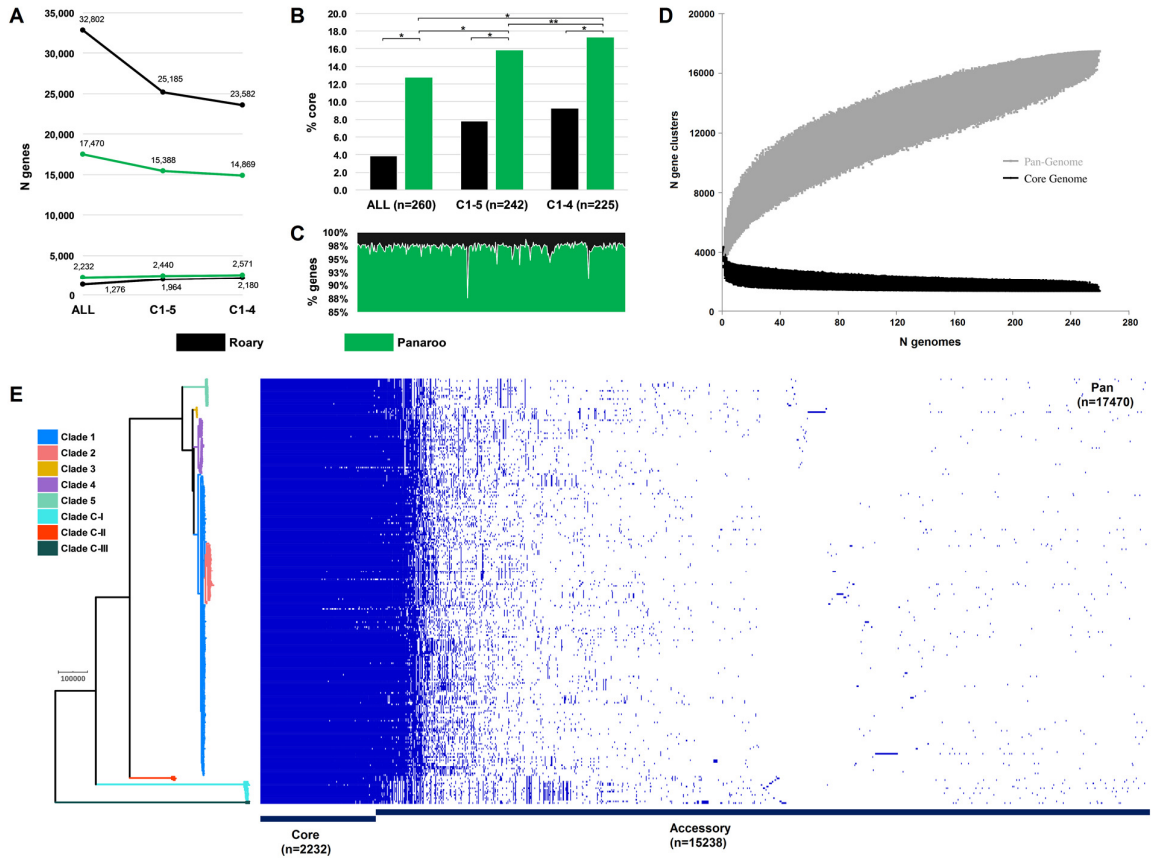


**Figure 6. *Clostridioides difficile* species pangenome. (A)** Pan and core genome estimates for all 260 STs, clades C1-4 (n=242 STs) and clades C1-5 (n=225 STs). **(B)** The difference in % core genome and pangenome sizes with Panaroo and Roary algorithms. (*) indicates $\chi^2$ $p < 0.00001$ and (**) indicates $\chi^2$ $p = 0.0008$. **(C)** The proportion of retained genes per genome after polishing Prokka-annotated genomes with Panaroo. **(D)** The total number of genes in the pan (grey) and core (black) genomes are plotted as a function of the number of genomes sequentially added (n=260). Following the definition of Tettelin *et al.*[28], the *C. difficile* species pangenome showed characteristics of an "open" pangenome. First, the pangenome increased in size exponentially with sampling of new genomes. At n=260, the pangenome exceeded more than double the average number of genes found in a single *C. difficile* genome (~3,700) and the curve was yet to reach a plateau or exponentially decay, indicating more sequenced strains are needed to capture the complete species gene repertoire. Second, the number of new 'strain-specific' genes did not converge to zero upon sequencing of additional strains, at n=260, an average of 27 new genes were contributed to the gene pool. Finally, according to Heap's Law, α values of ≤ 1 are representative of open pangenome. Rarefaction analysis of our pangenome curve using a power-law regression model based on Heap's Law[28] showed the pangenome was predicted to be open (*Bpan* ($\approx \alpha^{28}$) = 0.47, curve fit, $r^2$=0.999). **(E)** Presence absence variation (PAV) matrix for 260 *C. difficile* genomes is shown alongside a maximum-likelihood phylogeny built from a recombination-adjusted alignment of core genes from Panaroo (2,232 genes, 2,606,142 sites).

222 **Table 2 Major clade-specific gene clusters identified by pan-GWAS**

| Protein | Gene | Clade specificity | Functional insights |
|---|---|---|---|
| Ethanolamine kinase | *ETNK, EKI* | Unique to C-III and is in addition to the highly conserved *eut* cluster found in all lineages. Has a unique composition and includes six additional genes that are not present in the traditional CD630 *eut* operon or any other non-C-III strains. | An alternative process for the breakdown of ethanolamine and its utilisation as a source of reduced nitrogen and carbon. |
| Agmatinase | *speB* | | |
| 1-propanol dehydrogenase | *pduQ* | | |
| Ethanolamine utilization protein EutS | *eutS* | | |
| Ethanolamine utilization protein EutP | *eutP* | | |
| Ethanolamine ammonia-lyase large subunit | *eutB* | | |
| Ethanolamine ammonia-lyase small subunit | *eutC* | | |
| Ethanolamine utilization protein EutL | *eutL* | | |
| Ethanolamine utilization protein EutM | *eutM* | | |
| Acetaldehyde dehydrogenase | *E1.2.1.10* | | |
| Putative phosphotransacetylase | *K15024* | | |
| Ethanolamine utilization protein EutN | *eutN* | | |
| Ethanolamine utilization protein EutQ | *eutQ* | | |
| TfoX/Sxy family protein | - | | |
| Iron complex transport system permease protein | *ABC.FEV.P* | Unique to C-III | Multicomponent transport system with specificity for chelating heavy metal ions. |
| Iron complex transport system ATP-binding protein | *ABC.FEV.A* | | |
| Iron complex transport system substrate-binding protein | *ABC.FEV.S* | | |
| Hydrogenase nickel incorporation protein HypB | *hypB* | | |
| Putative ABC transport system ATP-binding protein | *yxdL* | | |
| Class I SAM-dependent methyltransferase | - | | |
| Peptide/nickel transport system substrate-binding protein | *ABC.PE.S* | | |
| Peptide/nickel transport system permease protein | *ABC.PE.P* | | |
| Peptide/nickel transport system permease protein | *ABC.PE.P1* | | |
| Peptide/nickel transport system ATP-binding protein | *ddpD* | | |
| Oligopeptide transport system ATP-binding protein | *oppF* | | |
| Class I SAM-dependent methyltransferase | - | | |
| Heterodisulfide reductase subunit D [EC:1.8.98.1] | *hdrD* | Unique to C-III and is in addition to the highly conserved spermidine uptake cluster found in all other lineages. | Alternative spermidine uptake processes which may play a role in stress response to nutrient limitation. The additional cluster has homologs in *Romboutsia, Paraclostridium* and *Paeniclostridium* spp. |
| CDP-L-myo-inositol myo-inositolphosphotransferase | *dipps* | | |
| Spermidine/putrescine transport system substrate-binding protein | *ABC.SP.S* | | |
| Spermidine/putrescine transport system permease protein | *ABC.SP.P1* | | |
| Spermidine/putrescine transport system permease protein | *ABC.SP.P* | | |
| Spermidine/putrescine transport system ATP-binding protein | *potA* | | |
| Sigma -54 dependent transcriptional regulator | *gfrR* | Present in all lineages except C-I. Cluster found in a different genomic position in C-III. | Mannose-type PTS system essential for utilisation of fructosamines such as fructoselysine and glucoselysine, abundant components of rotting fruit and vegetable matter. |
| Fructoselysine/glucoselysine PTS system EIIB component | *gfrB* | | |
| Mannose PTS system EIIA component | *manXa* | | |
| Fructoselysine/glucoselysine PTS system EIIC component | *gfrC* | | |
| Fructoselysine/glucoselysine PTS system EIID component | *gfrD* | | |
| SIS domain-containing protein | - | | |
| Fur family transcriptional regulator, ferric uptake regulator | *furB* | Unique to C-II and C5 | Associated with EDTA resistance in *E.coli*, helping the bacteria survive in Zn-depleted environment. |
| Zinc transport system substrate-binding protein | *znuA* | | |
| Fe-S-binding protein | *yeiR* | | |
| Rrf2 family transcriptional regulator | - | | |
| Putative signalling protein | - | Unique to C-I and C5 STs 163, 280, and 386 | In *E. coli*, AbgAB proteins enable uptake and cleavage of the folate catabolite *p*-aminobenzoyl-glutamate, allowing the bacterium to survive on exogenous sources of folic acid. |
| Aminobenzoyl-glutamate utilization protein B | *abgB* | | |
| MarR family transcriptional regulator | - | | |

223 **Cryptic clades CI-III possessed highly divergent toxin gene architecture.** Overall, 68.8%
224 (179/260) of STs harboured *tcdA* (toxin A) and/or *tcdB* (toxin B), indicating their ability to cause
225 CDI, while 67 STs (25.8%) harboured *cdtA*/*cdtB* (binary toxin). The most common genotype was
226 $A^+B^+CDT^-$ (113/187; 60.4%), followed by $A^+B^+CDT^+$ (49/187; 26.2%), $A^-B^+CDT^+$ (10/187; 5.3%),
227 $A^-B^-CDT^+$ (8/187; 4.3%) and $A^-B^+CDT^-$ (7/187; 3.7%). Toxin gene content varied across clades
228 (C1, 116/149, 77.9%; C2, 35/35, 100.0%; C3, 7/7, 100.0%; C4, 6/34, 17.6%; C5, 18/18, 100.0%;
229 C-I, 2/12, 16.7%; C-II, 1/3, 33.3%; C-III, 2/2, 100.0%) (**Fig. 7**).

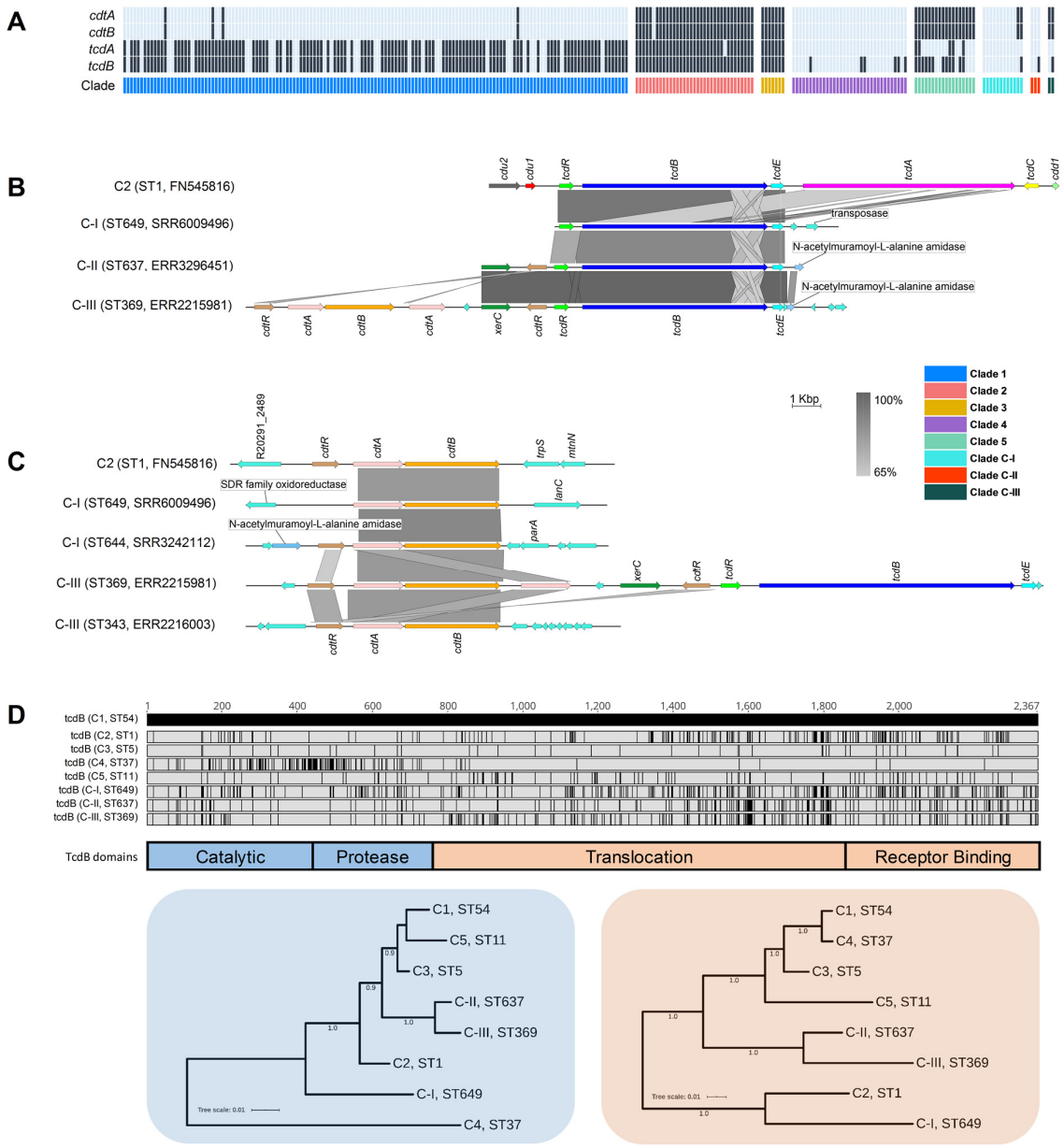


**Figure 7. Toxin gene analysis**. **(A)** Distribution of toxin genes across *C. difficile* clades (n=260 STs). Presence is indicated by black bars and absence by light blue bars. **(B)** Comparison of PaLoc architecture in the chromosome of strain R20291 (C2, ST1) and cognate chromosomal regions in genomes of cryptic STs 649 (C-I), 637 (C-II), and 369 (C-III). All three cryptic STs show atypical 'monotoxin' PaLoc structures, with the presence of syntenic *tcdR*, *tcdB*, and *tcdE*, and the absence of *tcdA*, *tcdC*, *cdd1* and *cdd2*. ST369 genome ERR2215981 shows colocalization of the PaLoc and CdtLoc, see below. **(C)** Comparison of CdtLoc architecture in the chromosome of strain R20291 (C2, ST1) and cognate chromosomal regions in genomes of cryptic STs 649/644 (C-I) and 343/369 (C-III). Several atypical CdtLoc features are observed; *cdtR* is absent in ST649, and an additional copy of *cdtA* is present in ST369, the latter comprising part of a CdtLoc co-located with the PaLoc. **(D)** Amino acid differences in TcdB among cryptic STs 649, 637, and 369 and reference strains from clades C1-5. Variations are shown as black lines relative to CD630 (C1, ST54). Phylogenies constructed from the catalytic and protease domains (in blue) and translocation and receptor-binding domains (in orange) of TcdB for the same eight STs included in (D). Scale bar shows the number of amino acid substitutions per site. Trees are mid-point rooted and supported by 500 bootstrap replicates.

Critically, at least one ST in each of clades C-I, C-II and C-III harboured divergent *tcdB* (89-94% identity to $tcdB_{R20291}$) and/or *cdtAB* alleles (60-71% identity to $cdtA_{R20291}$, 74-81% identity to $cdtB_{R20291}$). These genes were located on atypical and novel PaLoc and CdtLoc structures flanked by mediators of lateral gene transfer (**Fig. 7**). Sequence types 359, 360, 361 and 649 (C-I), 637 (C-II) and 369 (C-III) harboured 'monotoxin' PaLocs characterised by the presence of syntenic *tcdR*, *tcdB* and *tcdE*, and complete absence of *tcdA* and *tcdC*. In STs 360 and 361 (C-I), and 637 (C-II), a gene encoding an endolysin with predicted N-acetylmuramoyl-L-alanine amidase activity (*cwlH*) was found adjacent to the phage-derived holin gene *tcdE*.

Remarkably, a full CdtLoc was found upstream of the PaLoc in ST369 (C-III). This CdtLoc was unusual, characterised by the presence of *cdtB*, two copies of *cdtA*, two copies of *cdtR* and *xerC* encoding a site-specific tyrosine recombinase (**Fig. 7**). Both ST644 (C-I) and ST343 (C-III) were CdtLoc-positive but PaLoc-negative ($A^-B^-CDT^+$). In ST649 (C-I) *cdtR* was completely absent and, in ST343 (C-III), the entire CdtLoc was contained within the genome of a 56Kbp temperate bacteriophage termed ΦSemix9P1[29]. Toxin regulators TcdR and CdtR are highly conserved across clades C1-5[21]. In contrast, the CdtR of STs 644 (C-I), 343 (C-III) and 369 (C-III) shared only 46-54% amino acid identity (AAI) with CdtR of strain R20291 from clade 2 and ~40% AAI to each other. Similarly, the TcdR of ST 369 shared only 82.1% AAI compared to R20291 (**Supplementary Data**).

Compared to TcdB of R20291 ($TcdB_{R20291}$), the shared AAI for $TcdB_{ST649_C\text{-}I}$, $TcdB_{ST637_C\text{-}II}$ and $TcdB_{ST369_C\text{-}III}$ were 94.0%, 90.5% and 89.4%, respectively. This sequence heterogeneity was confirmed through the detection of five distinct *Hinc*II/*Acc*I digestion profiles of *tcdB* B1 fragments possibly reflecting novel toxinotypes (**Supplementary Data**). TcdB phylogenies identified clade C2 as the most recent common ancestor for $TcdB_{ST649_C\text{-}I}$ (**Fig. 7**). Phylogenetic subtyping analysis of the TcdB receptor-binding domain (RBD) showed the respective sequences in C-I, C-II and C-III clustered with *tcdB* alleles belonging to virulent C2 strains (**Supplementary Data**). Notably, the TcdB-RBD of ST649 (C-I) shared an AAI of 93.5% with TcdB-RBD allele type 8 belonging to hypervirulent STs 1 (RT027)[13] and 231 (RT251)[30]. Similarly, the closest match to *tcdB*-RBDs of ST637 (C-II) and ST369 (C-III) was allele type 10 (ST41, RT244)[31].

# Discussion

Through phylogenomic analysis of the largest and most diverse collection of *C. difficile* genomes to date, we identified major incoherence in *C. difficile* taxonomy, and provide new insight into intra-species diversity and evolution of pathogenicity in this major One Health pathogen.

Our analysis found high nucleotide identity (ANI > 97%) between *C. difficile* clades C1-4, indicating that strains from these four clades (comprising 560 known STs) belong to the same species. This is supported by our core genome and Bayesian analyses, which estimated the most recent common ancestor of *C. difficile* clades C1-4 existed ~1.61 mya. After this point, there appears to have been rapid population expansion into the four closely related extant clades described today, which include many of the most prevalent strains causing healthcare-associated CDI worldwide[11]. On the other hand, ANI between C5 and C1-4 is on the borderline of the accepted species threshold (95.9-96.2%) and their common ancestor existed 3.89 mya, over 2 Ma before C1-4 diverged. This degree of speciation likely reflects the unique ecology of C5 – a lineage comprising 33 known STs which is well established in non-human animal reservoirs worldwide and recently associated with CDI in the community setting[32]. We identified major taxonomic incoherence among the three cryptic clades and C1-5, evident by ANI values well below the species threshold (~91%, C-I; ~94%, C-II; and ~89%, C-III). Similar ANI value differences were seen between the cryptic clades themselves, indicating they are as divergent from each other as they are individually from C1-5. This extraordinary level of discontinuity is substantiated by our core genome and Bayesian analyses which estimated the common ancestors of clades C-I, C-II and C-III existed 13, 22 and 48 Ma, respectively, at least 9 to 45 Ma before the common ancestor of C1-5. For context, divergence dates for other pathogens range from 10 Ma (*Campylobacter coli* and *C. jejuni*)[33], 47 Ma (*Burkholderia pseudomallei* and *B. thailandensis*)[34] and 120 Ma (*Escherichia coli* and *Salmonella enterica*)[35]. Corresponding whole genome ANI values for these species are 86%, 94% and 82%, respectively (**Supplementary Data**).

295 Comparative ANI analysis of the cryptic clades with >5000 reference genomes across 21
296 phyla failed to provide a better match than *C. difficile* (89-94% ANI). Similarly, our revised ANI-
297 based taxonomy of the *Peptostreptococcaceae* placed clades C-I, C-II and C-III between *C. difficile*
298 and *C. mangenotii*, the latter sharing ~77% ANI. The rate of 16S rRNA divergence in bacteria is
299 estimated to be 1–2% per 50 Ma[35]. Contradicting our ANI and core genome data, 16S rRNA
300 sequences were highly conserved across all 8 clades. This indicates that in *C. difficile*, 16S rRNA
301 gene similarity correlates poorly with measures of genomic, phenotypic and ecological diversity, as
302 reported in other taxa such as *Streptomyces, Bacillus* and *Enterobacteriaceae*[36, 37]. Another interesting
303 observation is that C5 and the three cryptic clades had a high proportion (>90%) of MLST alleles that
304 were absent in other clades (**Supplementary Data**) suggesting minimal exchange of essential
305 housekeeping genes between these clades. Whether this reflects divergence or convergence of two
306 species, as seen in *Campylobacter*[38], is unknown. Taken together, these data strongly support the
307 reclassification of *C. difficile* clades C-I, C-II and C-III as novel independent *Clostridioides*
308 genomospecies. There have been similar genome-based reclassifications in *Bacillus*[39],
309 *Fusobacterium*[40] and *Burkholderia*[41]. Also, a recent Consensus Statement[42] argues that the genomics
310 and big data era necessitate easing of nomenclature rules to accommodate genome-based assignment
311 of species status to nonculturable bacteria and those without 'type material', as is the case with these
312 genomospecies.
313 The NCBI SRA was dominated by C1 and C2 strains, both in number and diversity. This
314 apparent bias reflects the research community's efforts to sequence the most prominent strains
315 causing CDI in regions with the highest-burden, e.g. ST 1 from humans in Europe and North America.
316 As such, there is a paucity of sequenced strains from diverse environmental sources, animal reservoirs
317 or regions associated with atypical phenotypes. Cultivation bias - a historical tendency to culture,
318 preserve and ultimately sequence *C. difficile* isolates that are concordant with expected phenotypic
319 criteria, comes at the expense of 'outliers' or intermediate phenotypes. Members of the cryptic clades
320 fit this criterion. They were first identified in 2012 but have been overlooked due to atypical toxin
321 architecture which may compromise diagnostic assays (discussed below). Our updated MLST
322 phylogeny shows as many as 55 STs across the three cryptic clades (C-I, n=25; C-II, n=9; C-III, n=21)
323 (**Fig. 2**). There remains a further dozen 'outliers' which could either fit within these new taxa or be
324 the first typed representative of additional genomospecies. The growing popularity of metagenomic
325 sequencing of animal and environmental microbiomes will certainly identify further diversity within
326 these taxa, including nonculturable strains[43, 44].
327 By analysing 260 STs across eight clades, we provide the most comprehensive pangenome
328 analysis of *C. difficile* to date. Importantly, we also show that the choice of algorithm significantly
329 affects pangenome estimation. The *C. difficile* pangenome was determined to be open (i.e. an
330 unlimited gene repertoire) and vast in scale (over 17000 genes), much larger than previous estimates
331 (~10000 genes) which mainly considered individual clonal lineages[16, 22]. Conversely, comprising just
332 12.8% of its genetic repertoire (2,232 genes), the core genome of *C. difficile* is remarkably small,
333 consistent with earlier WGS and microarray-based studies describing ultralow genome conservation
334 in *C. difficile*[11, 45]. Considering only C1-5, the pangenome reduced in size by 12% (2,082 genes);
335 another 519 genes were lost when considering only C1-4. These findings are consistent with our
336 taxonomic data, suggesting the cryptic clades, and to a lesser extent C5, contribute a significant
337 proportion of evolutionarily divergent and unique loci to the gene pool. A large open pangenome and
338 small core genome are synonymous with a sympatric lifestyle, characterised by cohabitation with,
339 and extensive gene transfer between, diverse communities of prokarya and archaea[46]. Indeed,
340 *C. difficile* shows a highly mosaic genome comprising many phages, plasmids and integrative and
341 conjugative elements[11], and has adapted to survival in multiple niches including the mammalian
342 gastrointestinal tract, water, soil and compost, and invertebrates[32].
343 Through a robust Pan-GWAS approach we identified loci that are enriched or unique in the
344 genomospecies. C-I strains were associated with the presence of transporter AbgB and absence of a
345 mannose-type phosphotransferase (PTS) system. In *E. coli*, AbgAB proteins allow it to survive on
346 exogenous sources of folate[47]. In many enteric species, the mannose-type PTS system is essential for

347 catabolism of fructosamines such as glucoselysine and fructoselysine, abundant components of
348 rotting fruit and vegetable matter[48]. C-II strains contained Zn transporter loci *znuA* and *yeiR*, in
349 addition to Zn transporter ZupT which is highly conserved across all eight *C. difficile* clades.
350 *S. enterica* and *E. coli* harbour both *znuA*/*yeiR* and ZupT loci, enabling survival in Zn-depleted
351 environments[49]. C-III strains were associated with major gene clusters encoding systems for
352 ethanolamine catabolism, heavy metal transport and spermidine uptake. The C-III *eut* gene cluster
353 encoded six additional kinases, transporters and transcription regulators absent from the highly
354 conserved *eut* operon found in other clades. Ethanolamine is a valuable source of carbon and/or
355 nitrogen for many bacteria, and *eut* gene mutations (in C1/C2) impact toxin production *in vivo*[50]. The
356 C-III metal transport gene cluster encoded a chelator of heavy metal ions and a multi-component
357 transport system with specificity for iron, nickel and glutathione. The conserved spermidine operon
358 found in all *C. difficile* clades is thought to play an important role in various stress responses including
359 during iron limitation[51]. The additional, divergent spermidine transporters found in C-III were similar
360 to regions in closely related genera *Romboutsia* and *Paeniclostridium* (data not shown). Together,
361 these data provide preliminary insights into the biology and ecology of the genomospecies. Most
362 differential loci identified were responsible for extra or alternate metabolic processes, some not
363 previously reported in *C. difficile*. It is therefore tempting to speculate that the evolution of alternate
364 biosynthesis pathways in these species reflects distinct ancestries and metabolic responses to evolving
365 within markedly different ecological niches.
366 This work demonstrates the presence of toxin genes on PaLoc and CdtLoc structures in all
367 three genomospecies, confirming their clinical relevance. Monotoxin PaLocs were characterised by
368 the presence of *tcdR*, *tcdB* and *tcdE*, the absence of *tcdA* and *tcdC*, and flanking by transposases and
369 recombinases which mediate LGT[20, 21, 52]. These findings support the notion that the classical bi-toxin
370 PaLoc common to clades C1-5 was derived by multiple independent acquisitions and stable fusion of
371 monotoxin PaLocs from ancestral Clostridia[52]. Moreover, the presence of syntenic PaLoc and CdtLoc
372 (in ST369, C-I), the latter featuring two copies of *cdtA* and *cdtR,* and a recombinase (*xerC*), further
373 support this PaLoc fusion hypothesis[52].
374 Bacteriophage holin and endolysin enzymes coordinate host cell lysis, phage release and toxin
375 secretion[53]. Monotoxin PaLocs comprising phage-derived holin (*tcdE*) and endolysin (*cwlH*) genes
376 were first described in C-I strains[52]. We have expanded this previous knowledge by demonstrating
377 that syntenic *tcdE* and *cwlH* are present within monotoxin PaLocs across all three genomospecies.
378 Moreover, since some strains contained *cwlH* but lacked toxin genes, this gene seems to be implicated
379 in toxin acquisition. These data, along with the detection of a complete and functional[29] CdtLoc
380 contained within ΦSemix9P1 in ST343 (C-III), further substantiate the role of phages in the evolution
381 of toxin loci in *C. difficile* and related Clostridia [53].
382 The CdtR and TcdR sequences of the new genomospecies are unique and further work is
383 needed to determine if these regulators display different mechanisms or efficiencies of toxin
384 expression[12]. The presence of dual copies of CdtR in ST369 (C-I) is intriguing, as analogous
385 duplications in PaLoc regulators have not been documented. One of these CdtR had a mutation at a
386 key phosphorylation site (Asp61→Asn61) and possibly shows either reduced wild-type activity or
387 non-functionality, as seen in ST11[54]. This might explain the presence of a second CdtR copy.
388 TcdB alone can induce host innate immune and inflammatory responses leading to intestinal
389 and systemic organ damage[55]. Our phylogenetic analysis shows TcdB sequences from the three
390 genomospecies are related to TcdB in Clade 2 members, specifically ST1 and ST41, both virulent
391 lineages associated with international CDI outbreaks[13, 31], and causing classical or variant
392 (*C. sordellii*-like) cytopathic effects, respectively[56]. It would be relevant to explore whether the
393 divergent PaLoc and CdtLoc regions confer differences in biological activity, as these may present
394 challenges for the development of effective broad-spectrum diagnostic assays, and vaccines. We have
395 previously demonstrated that common laboratory diagnostic assays may be challenged by changes in
396 the PaLoc of C-I strains[21]. The same might be true for monoclonal antibody-based treatments for CDI
397 such as bezlotoxumab, known to have distinct neutralizing activities against different TcdB
398 subtypes[57].

399 Our findings highlight major incongruence in *C. difficile* taxonomy, identify differential
400 patterns of diversity among major clades and advance understanding of the evolution of the PaLoc
401 and CdtLoc. While our analysis is limited solely to the genomic differences between *C. difficile*
402 clades, our data provide a robust genetic foundation for future studies to focus on the phenotypic,
403 ecological and epidemiological features of these interesting groups of strains, including defining the
404 biological consequences of clade-specific genes and pathogenic differences *in vitro* and *in vivo*.
405 Finally, our findings reinforce that the epidemiology of this important One Health pathogen is not
406 fully understood. Enhanced surveillance of CDI and WGS of new and emerging strains to better
407 inform the design of diagnostic tests and vaccines are key steps in combating the ongoing threat posed
408 by *C. difficile*.

409 # Methods

410 **Genome collection.** We retrieved the entire collection of *C. difficile* genomes (taxid ID 1496) held
411 at the NCBI Sequence Read Archive [https://www.ncbi.nlm.nih.gov/sra/]. The raw dataset (as of 1st
412 January 2020), comprised 12,621 genomes. After filtering for redundancy and Illumina paired-end
413 data (all platforms and read lengths), 12,304 genomes (97.5%) were available for analysis.

414 **Multi-locus sequence typing.** Sequence reads were interrogated for multi-locus sequence type (ST)
415 using SRST2 v0.1.8[58]. New alleles, STs and clade assignments were verified by submission of
416 assembled contigs to PubMLST [https://pubmlst.org/cdifficile/]. A species-wide phylogeny was
417 generated from 659 ST alleles sourced from PubMLST (dated 01-Jan-2020). Alleles were
418 concatenated in frame and aligned with MAFFT v7.304. A final neighbour-joining tree was generated
419 in MEGA v10[59] and annotated using iToL v4 [https://itol.embl.de/].

420 **Genome assembly and quality control**. Genomes were assembled, annotated and evaluated using a
421 pipeline comprising TrimGalore v0.6.5, SPAdes v3.6.043, Prokka v1.14.5, and QUAST v2.344[16].
422 Next, Kraken2 v2.0.8-beta[60] was used to screen for contamination and assign taxonomic labels to
423 reads and draft assemblies.

424 **Taxonomic analyses**. Species-wide genetic similarity was determined by computation of whole-
425 genome ANI for 260 STs. Both alignment-free and conventional alignment-based ANI approaches
426 were taken, implemented in FastANI[5] v1.3 and the Python module pyani[61] v0.2.9, respectively.
427 FastANI calculates ANI using a unique *k*-mer based alignment-free sequence mapping engine, whilst
428 pyani utilises two different classical alignment ANI algorithms based on BLAST+ (ANIb) and
429 MUMmer (ANIm). A 96% ANI cut-off was used to define species boundaries[4]. For taxonomic
430 placement, ANI was determined for divergent *C. difficile* genomes against two datasets comprising
431 (i) members of the *Peptostreptococcaceae* (n=25)[23], and (ii) the complete NCBI RefSeq database
432 (n=5895 genomes, https://www.ncbi.nlm.nih.gov/refseq/, accessed 14th Jan 2020). Finally,
433 comparative identity analysis of consensus 16S rRNA sequences for *C. mangenotii* type strain
434 DSM1289T[23] (accession FR733662.1) and representatives of each *C. difficile* clade was performed
435 using Clustal Omega https://www.ebi.ac.uk/Tools/msa/clustalo/.

436 **Estimates of clade and species divergence.** BactDating v1.0.1[62] was applied to the recombination-
437 corrected phylogeny produced by Gubbins (471,708 core-genome sites) with Markov chain Monte
438 Carlo (MCMC) chains of $10^7$ iterations sampled every $10^4$ iterations with a 50% burn-in. A strict
439 clock model was used with a rate of $2.5\times10^{-9}$ to $1.5\times10^{-8}$ substitutions per site per year, as previously
440 defined by He *et al.*[16] and Kumar *et al.*[27]. The effective sample sizes (ESS) were >200 for all estimated
441 parameters, and traces were inspected manually to ensure convergence. To provide an independent
442 estimate from BactDating, BEAST v1.10.4[63] was run on a recombination-filtered gap-free alignment
443 of 10,466 sites with MCMC chains of $5\times10^8$ iterations, with a $9\times10^{-7}$ burn-in, that were sampled
444 every $10^4$ iterations. The strict clock model described above was used in combination with the discrete
445 GTR gamma model of heterogeneity among sites and skyline population model. MCMC convergence
446 was verified with Tracer v1.7.1 and ESS for all estimated parameters were >150. For ease of

comparison, clade dating from both approaches were transposed onto a single MLST phylogeny. Tree files are available as **Supplementary Data** at http://doi.org/10.6084/m9.figshare.12471461.

**Pangenome analysis.** The 260 ST dataset was used for pangenome analysis with Panaroo v1.1.0[64] and Roary v3.6.0[65]. Panaroo was run with default thresholds for core assignment (98%) and blastP identity (95%). Roary was run with a default threshold for core assignment (99%) and two different thresholds for BlastP identity (95%, 90%). Sequence alignment of the final set of core genes (Panaroo; n=2,232 genes, 2,606,142 bp) was performed using MAFFT v7.304 and recombinative sites were filtered using Gubbins v7.304[66]. A recombinant adjusted alignment of 471,708 polymorphic sites was used to create a core genome phylogeny with RAxML v8.2.12 (GTR gamma model of among-site rate-heterogeneity), which was visualised alongside pangenome data in Phandango[67]. Pangenome dynamics were investigated with PanGP v1.0.1[16].

Scoary[68] v1.6.16 was used to identify genetic loci that were statistically associated with each clade via a Pangenome-Wide Association Study (pan-GWAS). The Panaroo-derived pangenome (n=17,470) was used as input for Scoary with the evolutionary clade of each genome depicted as a discrete binary trait. Scoary was run with 1,000 permutation replicates and genes were reported as significantly associated with a trait if they attained *p*-values (empirical, naïve and Benjamini-Hochberg-corrected) of ≤0.05, a sensitivity and specificity of > 99% and 97.5%, respectively, and were not annotated as "hypothetical proteins". All significantly associated genes were reannotated using prokka and BlastP and functional classification (KEGG orthology) was performed using the Koala suite of web-based annotation tools[69].

**Comparative analysis of toxin gene architecture**. The 260 ST genome dataset was screened for the presence of *tcdA*, *tcdB*, *cdtA* and *cdtB* using the Virulence Factors Database (VFDB) compiled within ABRicate v1.0 [https://github.com/tseemann/abricate]. Results were corroborated by screening raw reads against the VFDB using SRST2 v0.1.8[58]. Both approaches employed minimum coverage and identity thresholds of 90 and 75%, respectively. Comparative analysis of PaLoc and CdtLoc architecture was performed by mapping of reads with Bowtie2 v.2.4.1 to cognate regions in reference strain R20291 (ST1, FN545816). All PaLoc and CdtLoc loci investigated showed sufficient coverage for accurate annotation and structural inference. Genome comparisons were visualized using ACT and figures prepared with Easyfig[21]. MUSCLE-aligned TcdB sequences were visualized in Geneious v2020.1.2 and used to create trees in iToL v4.

**Statistical analyses.** All statistical analyses were performed using SPSS v26.0 (IBM, NY, USA). For pangenome analyses, Chi-squared test with Yate's correction was used to compare the proportion of core genes and a One-tailed Mann-Whitney U test was used to demonstrate the reduction of gene content per genome, with a p-value ≤ 0.05 considered statistically significant.

# References


1. Doolittle WF, Papke RT. Genomics and the bacterial species problem. *Genome Biol* **7**, 116 (2006).

2. Konstantinidis KT, Ramette A, Tiedje JM. The bacterial species definition in the genomic era. *Philos Trans R Soc Lond B Biol Sci* **361**, 1929-1940 (2006).

3. Wayne LG, *et al.* Report of the ad hoc committee on reconciliation of approaches to bacterial systematics. *Int J Syst Evol Microbiol* **37**, 463-464 (1987).

4. Ciufo S, *et al.* Using average nucleotide identity to improve taxonomic assignments in prokaryotic genomes at the NCBI. *Int J Syst Evol Microbiol* **68**, 2386-2392 (2018).

494 5. Jain C, Rodriguez RL, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis
495 of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun* **9**, 5114 (2018).
496
497 6. Richter M, Rossello-Mora R. Shifting the genomic gold standard for the prokaryotic species
498 definition. *Proc Natl Acad Sci U S A* **106**, 19126-19131 (2009).
499
500 7. Guh AY*, et al.* Trends in US burden of *Clostridioides difficile* infection and outcomes. *N Engl J*
501 *Med* **382**, 1320-1330 (2020).
502
503 8. CDC. Antibiotic resistance threats in the United States, 2013. Centers for Disease Control and
504 Prevention. Web citation: http://www.cdc.gov/drugresistance/threat-report-2013/. (2013).
505
506 9. CDC. Antibiotic resistance threats in the United States, 2019. Centers for Disease Control and
507 Prevention. Web citation: https://www.cdc.gov/drugresistance/biggest-threats.html., (2019).
508
509 10. Lim S, Knight D, Riley T. *Clostridium difficile* and One Health. *Clinical Microbiology and*
510 *Infection*, (2019).
511
512 11. Knight DR, Elliott B, Chang BJ, Perkins TT, Riley TV. Diversity and evolution in the genome
513 of *Clostridium difficile*. *Clin Microbiol Rev* **28**, 721-741 (2015).
514
515 12. Chandrasekaran R, Lacy DB. The role of toxins in *Clostridium difficile* infection. *FEMS*
516 *Microbiol Rev* **41**, 723-750 (2017).
517
518 13. He M*, et al.* Emergence and global spread of epidemic healthcare-associated *Clostridium*
519 *difficile*. *Nat Genet* **45**, 109-113 (2013).
520
521 14. Shaw HA*, et al.* The recent emergence of a highly related virulent *Clostridium difficile* clade
522 with unique characteristics. *Clin Microbiol Infect* **26**, 492-498 (2020).
523
524 15. Imwattana K*, et al. Clostridium difficile* ribotype 017 - characterization, evolution and
525 epidemiology of the dominant strain in Asia. *Emerg Microb Infect* **8**, 796-807 (2019).
526
527 16. Knight DR*, et al.* Evolutionary and genomic insights into *Clostridioides difficile* sequence type
528 11: a diverse, zoonotic and antimicrobial resistant lineage of global One Health importance. *MBio*
529 **10**, e00446-00419 (2019).
530
531 17. Dingle KE*, et al.* Evolutionary history of the *Clostridium difficile* pathogenicity locus. *Genome*
532 *Biol Evol* **6**, 36-52 (2014).
533
534 18. Didelot X*, et al.* Microevolutionary analysis of *Clostridium difficile* genomes to investigate
535 transmission. *Genome Biol* **13**, R118 (2012).
536
537 19. Janezic S, Potocnik M, Zidaric V, Rupnik M. Highly divergent *Clostridium difficile* strains
538 isolated from the environment. *PLoS One* **11**, e0167101 (2016).
539
540 20. Ramirez-Vargas G, Rodriguez C. Putative conjugative plasmids with *tcdB* and *cdtAB* genes in
541 *Clostridioides difficile*. *Clin Infect Dis* **26**, 2287-2290 (2020).
542
543 21. Ramírez-Vargas G*, et al.* Novel Clade CI *Clostridium difficile* strains escape diagnostic tests,
544 differ in pathogenicity potential and carry toxins on extrachromosomal elements. *Sci Rep* **8**, 1-
545 11 (2018).

546
547 22. Knight DR, Squire MM, Collins DA, Riley TV. Genome analysis of *Clostridium difficile* PCR
548 ribotype 014 lineage in Australian pigs and humans reveals a diverse genetic repertoire and
549 signatures of long-range interspecies transmission. *Front Microbiol* **7**, 2138 (2017).
550
551 23. Lawson PA, Citron DM, Tyrrell KL, Finegold SM. Reclassification of *Clostridium difficile* as
552 *Clostridioides difficile* (Hall and O'Toole 1935) Prevot 1938. *Anaerobe* **40**, 95-99 (2016).
553
554 24. Oren A, Rupnik M. *Clostridium difficile* and *Clostridioides difficile*: Two validly published and
555 correct names. *Anaerobe* **52**, 125-126 (2018).
556
557 25. Knetsch CW*, et al.* Comparative analysis of an expanded *Clostridium difficile* reference strain
558 collection reveals genetic diversity and evolution through six lineages. *Infect Genet Evol* **12**,
559 1577-1585 (2012).
560
561 26. He M*, et al.* Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc*
562 *Natl Acad Sci U S A* **107**, 7527-7532 (2010).
563
564 27. Kumar N*, et al.* Adaptation of host transmission cycle during *Clostridium difficile* speciation.
565 *Nat Genet* **51**, 1315-1320 (2019).
566
567 28. Tettelin H*, et al.* Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*:
568 implications for the microbial "pan-genome". *Proc Natl Acad Sci U S A* **102**, 13950-13955
569 (2005).
570
571 29. Riedel T*, et al.* A *Clostridioides difficile* bacteriophage genome encodes functional binary toxin-
572 associated genes. *J Biotechnol* **250**, 23-28 (2017).
573
574 30. Hong S, Knight DR, Chang B, Carman RJ, Riley TV. Phenotypic characterisation of *Clostridium*
575 *difficile* PCR ribotype 251, an emerging multi-locus sequence type clade 2 strain in Australia.
576 *Anaerobe* **60**, 102066 (2019).
577
578 31. Eyre DW*, et al.* Emergence and spread of predominantly community-onset *Clostridium difficile*
579 PCR ribotype 244 infection in Australia, 2010 to 2012. *Euro Surveill* **20**, 21059 (2015).
580
581 32. Knight DR, Riley TV. Genomic delineation of zoonotic origins of *Clostridium difficile*. *Front*
582 *Pub Health* **7**, 164 (2019).
583
584 33. Sheppard SK, Maiden MC. The evolution of *Campylobacter jejuni* and *Campylobacter coli*. *Cold*
585 *Spring Harb Perspect Biol* **7**, a018119 (2015).
586
587 34. Yu Y*, et al.* Genomic patterns of pathogen evolution revealed by comparison of *Burkholderia*
588 *pseudomallei*, the causative agent of melioidosis, to avirulent *Burkholderia thailandensis*. *BMC*
589 *Microbiol* **6**, 46 (2006).
590
591 35. Ochman H, Elwyn S, Moran NA. Calibrating bacterial evolution. *Proc Natl Acad Sci U S A* **96**,
592 12638-12643 (1999).
593
594 36. Janda JM, Abbott SL. 16S rRNA gene sequencing for bacterial identification in the diagnostic
595 laboratory: pluses, perils, and pitfalls. *J Clin Microbiol* **45**, 2761-2764 (2007).
596

37. Chevrette MG, Carlos-Shanley C, Louie KB, Bowen BP, Northen TR, Currie CR. Taxonomic and metabolic incongruence in the ancient genus *Streptomyces*. *Front Microbiol* **10**, 2170 (2019).

38. Sheppard SK, McCarthy ND, Falush D, Maiden MC. Convergence of *Campylobacter* species: implications for bacterial evolution. *Science* **320**, 237-239 (2008).

39. Liu Y, Lai QL, Shao ZZ. Genome analysis-based reclassification of *Bacillus weihenstephanensis* as a later heterotypic synonym of *Bacillus mycoides*. *Int J Syst Evol Microbiol* **68**, 106-112 (2018).

40. Kook JK*, et al.* Genome-based reclassification of *Fusobacterium nucleatum* subspecies at the species level. *Curr Microbiol* **74**, 1137-1147 (2017).

41. Loveridge EJ*, et al.* Reclassification of the specialized metabolite producer *Pseudomonas mesoacidophila* ATCC 31433 as a member of the *Burkholderia cepacia* complex. *J Bacteriol* **199**, e00125-00117 (2017).

42. Murray AE*, et al.* Roadmap for naming uncultivated Archaea and Bacteria. *Nat Microbiol* **5**, 987-994 (2020).

43. Stewart RD*, et al.* Assembly of 913 microbial genomes from metagenomic sequencing of the cow rumen. *Nat Commun* **9**, 1-11 (2018).

44. Lu X*, et al.* Bacterial pathogens and community composition in advanced sewage treatment systems revealed by metagenomics analysis based on high-throughput sequencing. *PLoS One* **10**, e0125549 (2015).

45. Scaria J, Ponnala L, Janvilisri T, Yan W, Mueller LA, Chang YF. Analysis of ultra low genome conservation in *Clostridium difficile*. *PLoS One* **5**, e15147 (2010).

46. Medini D, Donati C, Tettelin H, Masignani V, Rappuoli R. The microbial pan-genome. *Curr Opin Genet Dev* **15**, 589-594 (2005).

47. Carter EL, Jager L, Gardner L, Hall CC, Willis S, Green JM. *Escherichia coli* abg genes enable uptake and cleavage of the folate catabolite p-aminobenzoyl-glutamate. *J Bacteriol* **189**, 3329-3334 (2007).

48. Miller KA, Phillips RS, Kilgore PB, Smith GL, Hoover TR. A mannose family phosphotransferase system permease and associated enzymes are required for utilization of fructoselysine and glucoselysine in *Salmonella enterica* serovar Typhimurium. *J Bacteriol* **197**, 2831-2839 (2015).

49. Sabri M, Houle S, Dozois CM. Roles of the extraintestinal pathogenic *Escherichia coli* ZnuACB and ZupT zinc transporters during urinary tract infection. *Infect Immun* **77**, 1155-1164 (2009).

50. Nawrocki KL, Wetzel D, Jones JB, Woods EC, McBride SM. Ethanolamine is a valuable nutrient source that impacts *Clostridium difficile* pathogenesis. *Environ Microbiol* **20**, 1419-1435 (2018).

51. Berges M*, et al.* Iron regulation in *Clostridioides difficile*. *Front Microbiol* **9**, 3183 (2018).

52. Monot M*, et al. Clostridium difficile*: new insights into the evolution of the pathogenicity locus. *Sci Rep* **5**, 15023 (2015).

53. Fortier LC. Bacteriophages contribute to shaping *Clostridioides* (*Clostridium*) *difficile* species. *Front Microbiol* **9**, 2033 (2018).

54. Bilverstone TW, Minton NP, Kuehne SA. Phosphorylation and functionality of CdtR in *Clostridium difficile*. *Anaerobe* **58**, 103-109 (2019).

55. Carter GP, *et al.* Defining the roles of TcdA and TcdB in localized gastrointestinal disease, systemic organ damage, and the host response during *Clostridium difficile* infections. *MBio* **6**, e00551 (2015).

56. Lanis JM, Barua S, Ballard JD. Variations in TcdB activity and the hypervirulence of emerging strains of *Clostridium difficile*. *PLoS Pathog* **6**, e1001061 (2010).

57. Shen E, *et al.* Subtyping analysis reveals new variants and accelerated evolution of *Clostridioides difficile* toxin B. *Commun Biol* **3**, 1-8 (2020).

58. Inouye M, *et al.* SRST2: rapid genomic surveillance for public health and hospital microbiology labs. *Genome Med* **6**, 90 (2014).

59. Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol* **35**, 1547-1549 (2018).

60. Wood DE, Lu J, Langmead B. Improved metagenomic analysis with Kraken 2. *Genome Biol* **20**, 257 (2019).

61. Pritchard L, Glover RH, Humphris S, Elphinstone JG, Toth IK. Genomics and taxonomy in diagnostics for food security: soft-rotting enterobacterial plant pathogens. *Anal Methods* **8**, 12-24 (2016).

62. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res* **46**, e134-e134 (2018).

63. Drummond AJ, Rambaut A. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol Biol* **7**, 214 (2007).

64. Tonkin-Hill G, *et al.* Producing polished prokaryotic pangenomes with the panaroo pipeline. *Genome Biol* **21**, 180 (2020).

65. Page AJ, *et al.* Roary: rapid large-scale prokaryote pan genome analysis. *Bioinformatics* **31**, 3691-3693 (2015).

66. Croucher NJ, *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15 (2015).

67. Hadfield J, Croucher NJ, Goater RJ, Abudahab K, Aanensen DM, Harris SR. Phandango: an interactive viewer for bacterial population genomics. *Bioinformatics* **34**, 292-293 (2018).

68. Brynildsrud O, Bohlin J, Scheffer L, Eldholm V. Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary. *Genome Biol* **17**, 238 (2016).

700 69. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG tools for functional
701 characterization of genome and metagenome sequences. *J Mol Biol* **428**, 726-731 (2016).

## 702 Author contributions

703 D.R.K., K.I., D.W.E., and T.V.R. designed the study. D.R.K., K.I., C.R., B.K., E.G.A., and K.E.D.
704 performed experimental work. D.R.K., K.I., C.R., B.K., E.G.A., D.P.S., X.D., K.E.D., D.W.E., C.R.,
705 and T.V.R. analysed data and drafted the manuscript. All authors edited and approved the final
706 version of the manuscript. The corresponding author had full access to all the data in the study and
707 had final responsibility for the decision to submit for publication.

## 708 Acknowledgements

709 This work was supported, in part, by funding from The Raine Medical Research Foundation
710 (RPG002-19) and a Fellowship from the National Health and Medical Research Council
711 (APP1138257) awarded to D.R.K. K.I. is a recipient of the Mahidol Scholarship from Mahidol
712 University, Thailand. This work was also supported by EULac project 'Genomic Epidemiology of
713 *Clostridium difficile* in Latin America (T020076)" and by the Millennium Science Initiative of the
714 Ministry of Economy, Development and Tourism of Chile, grant 'Nucleus in the Biology of Intestinal
715 Microbiota" to D.P.S. This research used the facilities and services of the Pawsey Supercomputing
716 Centre [Perth, Western Australia] and the Australian Genome Research Facility [Melbourne,
717 Victoria].

## 718 Competing Interests

719 DWE declares lecture fees from Gilead, outside the submitted work. No other author has a conflict
720 of interest to declare.

## 721 Additional information

722 Supplementary Data is available at http://doi.org/10.6084/m9.figshare.12471461