Research Article

Phylogenomic analysis of the collagen-like BclA proteins in *Clostridioides difficile*.

Francisca Cid-Rojas[1,4,#], Enzo Guerrero-Araya[1,#], Christian Brito-Silva[1], Marjorie Pizarro-Guajardo[1,4], César Rodríguez-Sanchéz[1,5], Daniel Paredes-Sabja[1,6,*]

[1]ANID – Millennium Science Initiative Program – Millennium Nucleus in the Biology of the Intestinal Microbiota, Santiago, Chile.

[4]Department of Biology, Texas A&M University, College Station, Texas, USA

[5] Faculty of Microbiology & CIET, University of Costa Rica, San José, Costa Rica.

[6]Department of Biology, Texas A&M University, College Station, Texas, USA, dparedes-sabja@bio.tamu.edu

 # Contributed equally to this work.

Running Title: BclA variability in *C. difficile*

Key Words: *C. difficile* spores, collagen-like proteins, BclA, *C. difficile* clades, exosporium, pseudogenization.

*Corresponding Author. Dr. Daniel Paredes-Sabja, Department of Biology, Texas A&M University, College Station, TX, 77843, USA. E-mail: dparedes-sabja@bio.tamu.edu

Abstract

Clostridioides difficile is a Gram-positive, anaerobic, spore-forming bacterium and a major nosocomial pathogen, notable for its high genetic diversity. *C. difficile* has been classified into five classical phylogenetic clades (C1 to C5) and five cryptic clades (C-I to C-V), reflecting its extensive genome plasticity. In addition, *C. difficile* spores are considered essential for the onset, persistence and transmission of the disease, and their exosporium layer has hair-like projections formed by the BclA-family of proteins (BclA1, BclA2 and BclA3). Previous work in C1 and C2 strains has demonstrated that the absence of these proteins affects spore germination, pathogenesis, persistence, and recurrence of CDI. Nevertheless, the conservation of BclAs across different *C. difficile* clades remains unclear. In this work, genomics analysis in more than 25,000 *C. difficile* genomes revealed that the prevalence and variability of the BclAs was not conserved across classical and cryptic clades. The most represented clade on the dataset was C1, where roughly 50% of the genomes possessed all three *bclA* genes. Pseudogenization of *bclA1* or all *bclAs* was observed in C2 and C3, respectively. Additionally, the absence of *bclA1* in C4 and of both *bclA1* & *bclA3* in C5, further demonstrates the divergence of BclAs among *C. difficile* clades. Subsequent analysis revealed high variability in the central collagen-like region (CLR) of all three BclAs, and a highly conserved *bclA1* pseudogenization event in most members of C2. Overall, the extensive variability of *bclA*, attributed to the CLR, prevalence of a *bclA1* pseudogenization and complete absence of *bclAs* in certain clades, is likely to impact spore morphogenesis and pathogenesis across different *C. difficile* clades.

50 **Introduction**

51

52 *C. difficile*, a ubiquitous anaerobic spore-producing microbe, is frequently implicated in nosocomial or
53 antimicrobial-associated infections, collectively known as *C. difficile* infections (CDI) [1]. This
54 environmentally persistent pathogen is capable of causing a wide range of symptoms, from mild diarrhea
55 to severe pseudomembranous colitis [2]. As a leading cause of healthcare-associated infections, *C. difficile*
56 poses a significant burden on global health, with incidence rates ranging from 2 to 8 cases per 10,000
57 patient-days [3, 4]. Conventional treatment approaches for CDI involve antibiotics, such as vancomycin,
58 metronidazole, and fidaxomicin [5]. However, despite these interventions, approximately 15 to 35% of CDI
59 cases experience recurrence (R-CDI) [6].

60

61 *C. difficile*'s infection lifecycle begins with the germination of endogenous or recently acquired C. *difficile*
62 spores [7]. Two main virulence factors, the toxins TcdA and TcdB, are primarily responsible for the clinical
63 manifestations of *Clostridioides difficile* infection (CDI) [8, 9]. These toxins monoglucosylate host Rho
64 family GTPases, leading to downstream effects such as tight junction disruption, epithelial cell
65 detachment, and compromised intestinal barrier function[8, 9]. Additionally, toxin exposure prompts
66 intestinal epithelial cells to secrete pro-inflammatory cytokines, which attract neutrophils and initiate acute
67 mucosal inflammation[8, 9]. This inflammatory response further activates immune cells, intensifying tissue
68 damage.

69

70 *C. difficile* spores are the main virulence factor related with disease recurrence [10, 11]. During CDI, *C.*
71 *difficile* begins a sporulation cycle leading to the formation of newly dormant spores[10], that interact with
72 host surfaces[12-15], and persist to cause R-CDI[16]. Essential in these interactions are the hair-like projections
73 on the spore surface, where the collagen-like BclA3 protein is essential for their formation[16]. Two
74 additional BclA orthologues (BclA1 and BclA2) are encoded in some *C. difficile* strains. Work in strains
75 630 (clade 1) and R20291 (clade 2), shows their implication in spore-host interactions, colonization and
76 recurrence of CDI [16, 17].

77

78 *C. difficile* has tremendous genetic diversity [18, 19], where its core genome represents only 10-20 % of its
79 pan-genome. This likely contributes to *C. difficile* adaptation to a wide host-range and persist in
80 environmental reservoirs[20]. This diversity is evidenced in the five distinct phylogenetic classical clades

81 and five cryptic clades, based on multi-locus sequencing and core-genome phylogeny[18]. While Clade 1 is
82 the most represented clade, with the broadest geographic distribution [3, 21], clade 2 includes epidemically
83 relevant strains of ribotype 027 (RT027) lineage and mainly associated with hospital outbreaks [22-24]. Clade
84 3 are uncommon yet phenotypically share similarities with Clade 2 strains[25]. Clade 4 includes strains that
85 are often clindamycin- and fluoroquinolone-resistant, and associated with outbreaks in Asia, North
86 America and Europe[26]. Finally, Clade 5 is the most genetically distant clade from all classical clades and
87 based on its 96% average nucleotide identity (ANI) to all other four clades, it is hypothesized to be an
88 early predecessor of all four clades[27]. However, several genomospecies, with ANI values lower than 95%
89 to all 5 classical clades and with highly divergent toxin gene architecture have been identified and defined
90 as cryptic clades (CI to V)[3, 18, 28, 29].
91
92 The outermost layer of the spores, called exosporium, possess surface-proteins resembling hair-like
93 projections[11, 30]. Transmission electron microscopy (TEM) of the spores has revealed that exosporium
94 morphology seems to be strain-dependent, as spores from the epidemically relevant strains (including
95 R20291) have a hair-like nap, meanwhile 630, a strain widely used in scientific research, has a compact
96 exosporium layer without hair like structures[11]. These hairs resemble those present on spores from
97 members of the *B. cereus* group (*B. anthracis*), which have been determined to be formed by BclA
98 glycoproteins[31, 32]. Strong *et al.*[33] identified three *bclA* homologues on the genome of *C. difficile*; *bclA1,*
99 *bclA2* & *bclA3*
100
101 Bacterial proteins like *B. anthracis* BclA & BclB and *Streptococcus pyogenes* (*S. pyogenes*) Scl1 & Scl2,
102 have been thoroughly characterized previously[31, 32, 34-39]. These proteins were encountered to be displayed
103 on the outermost layer of their respective bacterial morphotype with a central triple-helix rod shape
104 structure composed of canonical GXY collagen repeats[40] and a globular end terminal domain[32, 36, 38, 39],
105 thus they are referred to as collagen-like proteins. Although they differ in length, the domain distribution,
106 N-terminal domain (NTD), collagen-like region (CLR), and C-terminal domain (CTD), is the same as *C.*
107 *difficile* exosporium BclA proteins (**Fig. 1A**).
108
109 Despite the wealth of genomic data for *C. difficile*, there is a lack of knowledge on the diversity of spore
110 surface proteins implicated in pathogenesis. Therefore, in this work we explored BclA prevalence and
111 variability in more than 25,000 *C. difficile* genomes and used to generate a core-genome phylogenetic

112 tree, where four new clades were introduced, representing the complex and changing taxonomical
113 organization of this pathogen; the tree was divided into classical clades (C1 to C7) and cryptic clades (C-
114 I to C-VII). Subsequent *bclA* locus search and analysis revealed a high variability of the central collagen-
115 like region (CLR) of the proteins, indistinctively of the clades to which they belong, and the finding of a
116 conserved *bclA1* pseudogenization event present across most members of C2, which was not observed for
117 *bclA2* or *bclA3.* Complementation with a full length *bclA1* from 630 strain (C1) to R20291 strain, which
118 is a member of C2, resulted in an increase of exosporium and hair-like projection's length. Overall, the
119 extensive variability of *bclA*, mainly attributed to their CLR, prevalence of a *bclA1* pseudogenization and
120 even the complete absence of *bclAs* in certain clades.
121
122

Materials and Methods

Genome collection

The complete dataset from Enterobase was obtained, which included 25,145 carefully selected *C. difficile* genomes, as well as diverse metadata, with a cutoff date of September 30, 2022. This dataset is a valuable resource that encompasses information on a wide variety of strains and conditions. In addition to the data provided by Enterobase, 21 additional genomes were incorporated from suspected cryptic isolates, with the aim of further expanding the diversity and scope of the analysis.

Genomes assemble

Genomes were assembled, annotated, and evaluated using a pipeline [41] comprising TrimGalore v0.6.5, SPAdes v3.6.043, Prokka v1.14.5, and QUAST v2.344. Next, Kraken2 v2.0.8-beta was used to screen for contamination and assign taxonomic labels to reads and draft assemblies.

Multi-locus sequence typing

The assembled genomes underwent a thorough analysis for Multi-locus sequence typing (MLST) utilizing FastMLST v0.0.15 [42] to identify any significant genetic variations. During this process, new alleles and sequence types (STs) were detected and subsequently verified by submitting the assembled contigs to the PubMLST database (https://pubmlst.org/cdifficile/) [43], a widely recognized public resource for molecular typing and microbial genome diversity research

Core-genome genes and recombination-free phylogeny

To identify core-genome genes in the analysis, DIAMOND blastp [44] search was employed, coupled with a bidirectional best hit approach, using the proteins predicted by Prodigal [45] from both isolated and predicted proteomes of *C. difficile* 630 as query sequences. A gene was designated as a core gene if it was present in at least 99% of the isolates. Following the identification of core genes, MAFFT [46] was used to align each gene individually, and subsequently concatenate the alignments to form a comprehensive dataset.

To account for the potential influence of recombination events on the analysis, it was applied ClonalFrameML [47] to remove any signs of recombination from the concatenated alignment. With a

recombination-free alignment at hand, a phylogenetic tree was constructed using IQ-tree [48], implementing 1000 ultrafast bootstrap replicates to assess the robustness of the inferred tree topology.

Upon completion of the phylogenetic tree, final clade delimitations were stablished, which allowed to group *C. difficile* isolates into distinct evolutionary lineages. This comprehensive approach to genome analysis and phylogenetic tree construction provides valuable insights into the genetic relationships and evolutionary history of *C. difficile* isolates in the dataset used.

3D structural inference

Alphafold3[49] was employed, a state-of-the-art protein folding prediction algorithm, to determine the three-dimensional structure of the proteins under investigation. The Alphafold3 system utilizes machine learning techniques to predict protein structures with remarkable accuracy, providing valuable insights into protein function and molecular interactions.

Predictions were performed with 24 iterative cycles (seeds). Running multiple seeds of the model helps to refine the prediction at each cycle, enhancing the overall reliability and accuracy of the predicted structures. The highest-ranking predicted structure (rank 0), that had the lowest predicted alignment error (PAE) was selected for further analysis. The predicted alignment error provides a quantitative measure of the accuracy of the predicted structure compared to the native structure.

Sequence alignment

Clustal Omega [50], a widely recognized and efficient tool for generating multiple sequence alignments, was used to align the sequences of interest. Clustal Omega is an advanced algorithm that excels in aligning both nucleotide and amino acid sequences, making it a popular choice for comparing homologous sequences and identifying conserved regions, which can provide crucial insights into protein function, structure, and evolutionary relationships. To perform the multiple sequence alignments, default parameters provided by the Clustal Omega software were used.

EMBOSS Needle, was used to perform pairwise sequence alignment. It creates an optimal global alignment using the Needleman-Wunsch algorithm. Default parameters provided by the software were used.

185

186 **Extraction and distribution of BclAs proteins**

187 To extract the sequences of the three BclAs of interest, each genome in the dataset was first annotated
188 using Prodigal [45], a widely used and accurate gene prediction tool designed for microbial genomes. This
189 allowed to identify the coding sequences within each genome and obtain a comprehensive set of gene
190 annotations for further analysis. Following the genome annotation step, a search for the upstream and
191 downstream genes of each BclA within the annotated genomes using DIAMOND [44] was performed. By
192 locating these flanking genes, it was possible to define the genomic region encompassing each BclA.
193 Subsequently, these regions were extracted for further investigation and analysis. To ensure accuracy and
194 precision in the extraction process, this final step was performed manually. Given the large number of
195 isolates in the dataset, the focus was set on single nucleotide polymorphisms (SNPs) present in the
196 alignment for constructing the core-genome phylogenetic tree. To achieve this, the program employed
197 was RapidNJ (https://github.com/somme89/rapidNJ), a fast and efficient algorithm for constructing
198 phylogenetic trees based on the Neighbor-Joining method. To minimize redundancy in the dataset and
199 improve computational efficiency, deduplication using CD-HIT [51] was performed, a widely utilized
200 clustering tool that identifies and removes duplicated sequences based on sequence similarity thresholds.
201 This step ensured that the dataset contained only unique sequences, enabling more accurate and
202 meaningful downstream analyses. Lastly, custom Python scripts were used to generate schematic
203 representations of the results. These visualizations provided a clear and concise representation of the
204 findings, facilitating the interpretation and communication of the insights gleaned from the study.

205

206 ## Results

207

208 **Structural modeling of *C. difficile* BclA proteins.**

209 *C. difficile* encodes at least three orthologues of the BclA family of collagen-like proteins[52]. *C. difficile*
210 strains 630 (Ribotype 012, clade 1) and R20291 (Ribotype 027, clade 2), despite belonging to different
211 clades, they possess similar BclA proteins and share almost identical genetic context. Amino acid
212 sequence comparison reveals that BclA2 of 630 and R20291 have ~ 95 % identity (**Fig. S3, S4**), while
213 BclA3 shares ~ 87 % identity (**Fig. S5, S6**). However, in the case of BclA1, identity dropped to ~ 51 %,
214 however reaches 97 % upon comparing only the NTDs (**Table S1**). This is due to the pseudogenization
215 of *bclA1* in R20291 strain by two nonsense mutations at: i) nucleotide positions A145T, resulting in an
216 early-stop codons that lead to formation of a small 48 amino acid NTD polypeptide; and ii) at nucleotide
217 position C739T which yields an additional stop codon (**Fig. 1A, S1, S2**). When the alignment was
218 performed by individual BclA domains, gap-free alignment regions revealed more than 93% identity was
219 observed for NTD and CTDs of BclA2 and BclA3, except for the CLR of BclA3 with 77% identity (**Table**
220 **S1**). As to better compare both BclA1 from 630 and R20291 strain from a protein level in further analysis,
221 and since $bclA1_{R20291}$ is pseudogenized but the downstream DNA sequence follows the same structure as
222 a canonical *bclA1* (**Fig. S1**), the two nonsense mutations were repaired to obtain a full length repaired
223 $BclA1_{R20291}$ ($rBclA1_{R20291}$). Overall, BclAs of both strains were alike with the CLR domain having the
224 least similarity.

225

226 Given the differences in BclA1 and similarities in BclA2 and BclA3 between strains 630 and R20291,
227 AlphaFold3 (AF3) [53] was used to create structural predictions of full-lengths (fl) $BclA1_{630}$, $rBclA1_{R20291}$,
228 $BclA2_{R20291}$ and $BclA3_{R20291}$. Initial monomeric predictions for all fl-BclAs show low confidence models
229 for almost the entire protein. Upon modeling the individual domains, monomeric CLR of all four proteins
230 proved to be disorganized with very low predicted confidence. However, monomeric CTDs of BclA2 and
231 BclA3, except for $BclA1_{630}$ and $rBclA1_{R20291}$, were confidently predicted as globular structures with low
232 expected positional error. Interestingly, the NTDs of $BclA1_{630}$ and $rBclA1_{R20291}$ were predicted to be
233 globular domains, while the NTDs of BclA2 and BclA3 had not predicted secondary structure (**Fig. S7A,**
234 **S7B - S10A)**. Among these NTDs, $rBclA1_{R20291}$ was predicted with higher confidence than that of
235 $BclA1_{630}$, likely due to the longer CLR domain in $BclA1_{630}$ that could be affecting AF3 predictions (**Fig.**

236 **S7A & S7B**). Overall, monomeric predictions of all BclA were confident mainly at the CTD, while failing
237 to accurately predict NTD and CLR.
238
239 Collagen proteins have a globular trimerization domain at one of their terminal domains which together
240 with their intrinsic GXY repeats, are driving forces for the formation of triple-helix structures[49, 54],
241 suggesting that all three BclA are likely to form stable predictable trimers. Strikingly, although
242 homotrimers were predicted for all fl-BclAs, the confidence of these interactions was limited mainly to
243 the CTDs, where BclA2 had the highest predicted homotrimerization confidence, forming a tight CTD-
244 protein complex (**Fig. S8B**), followed by BclA3 (**Fig. S9B**). By contrast, both $BclA1_{630}$ and $rBclA1_{R20291}$
245 had an apparent low confidence homotrimerization prediction (**Fig. S7C & S7D**). Unfortunately, AF3 was
246 unable to predict the triple-helix structure in all *C. difficile* the collagen-like BclA proteins, likely due to
247 limitations of the model.
248
249 The shortest collagen-like proteins that have structural properties of collagens are synthetic collagen
250 mimetic peptides (CMPs), which have at least 6 GXY repeats that can self-assemble into triple-helix
251 structures[55, 56]. Therefore, a truncated CLR with 10 GXY repeats was created in all three BclA to evaluate
252 their predictability using AF3. The predicted structures of 10 GXY BclAs improved compared to full-
253 length BclAs. Prediction of 10 GXY versions of BclA2 and BclA3 revealed a more accurate organization
254 of BclA trimers, comprising of an N-terminal domain, a truncated central collagen-like region, and a C-
255 terminal domain (**Fig. S10**). However, for trimerization of $BclA1_{630}$, although prediction of truncated
256 BclA1 improved compared to full length, trimerization of the CTD was not as confident as for truncated
257 versions of BclA2 and BclA3 (**Fig. S10**). Altogether, structural predictions strongly support that the CTDs
258 of BclA2 and BclA3 tend to interact and are likely trimerization domains. By contrast, BclA1 trimerization
259 had no confident prediction, even in the truncated BclA1 versions.
260
261 Collagen molecules bind to different types of collagens forming heterotrimeric proteins[49, 57, 58]. Therefore,
262 to test the hypothesis that *C. difficile* BclA orthologues could form heterotrimers, AF3 predictions were
263 employed to predict the structure of trimeric complexes composed of different BclA proteins.
264 Combination of distinct BclA sequences were input as separate chains, allowing AF3 to model potential
265 inter-chain interactions and assess compatibility in forming stable trimeric assemblies. Resulting
266 predictions revealed no expected interactions between the CTD of different BclA, even though most

267 molecules were predicted with high confidence as independent units: i) CTD_BclA2-BclA3$_{R20291}$ (**Fig.**
268 **S11**), and ii) CTD_BclA1-BclA2-BclA3$_{630}$ (**Fig. S12**). These results support the notion that *C. difficile*
269 BclAs do not form heterotrimeric protein complexes.
270
271 Further analysis of CTDs, demonstrates that BclA1$_{630}$, BclA2 and BclA3 is formed by total of 12, 9 and
272 14 predicted beta sheets, while only CTD of BclA3 has an alpha helix in between beta-sheets 2 and 3 (**Fig.**
273 **1B, 1D, 1F**). While they all form a homotrimer barrel-type structure predicted with high confidence,
274 BclA1$_{630}$ and BclA2 have a ball-like shape, and BclA3 a trapezoid-like shape that, when viewed from
275 above, forms an isosceles triangle. Interestingly, the areas of lowest confidence in the inference of the
276 three-dimensional structure are in the coils at the tip of each monomer conforming the structures (**Fig. 1C,**
277 **1E, 1G, 1H**). These structural differences suggest that BclA1, BclA2, and BclA3 are homologues of each
278 other and may have different functions in the exosporium.
279
280 **Structural similarities between *C. difficile* and *B. anthracis* BclA proteins.**
281 To investigate the relationship between *C. difficile's* BclAs and *B. anthracis's* collagen-like proteins (BclA
282 and BclB), AF3 was used to model the protein's CTD_homotrimers of both bacteria. *C. difficile*
283 BclA_CTD domains were independently superimposed to *B. anthracis* BclA_CTD and BclB_CTD using
284 ChimeraX (**Fig. 2A**). Calculation of RMSD (Root Mean Square Deviation) was used to quantify atomic
285 differences between the superimposed proteins. Both, BclA1_CTD$_{630}$ and BclA2_CTD$_{R20291}$ had the best
286 hit with BclA_CTD of *B. anthracis*, with an RMSD of 0.956 and 1.078, respectively (**Fig. 2B and 2C**).
287 On the other hand, BclA3_CTD$_{R20291}$ showed a high structural similarity with BclB_CTD *B. anthracis*,
288 with an RMSD of 0.745 (**Fig. 2D**). Additionally, low sequence identity was observed between these
289 domain (15.8-19.5%) and a similarity between 23.9-35.6% (**Fig. S13**). These results suggest that despite
290 the low identity between the CTDs of *C. difficile* and *B. anthracis*, they share comparable predicted
291 structures.
292
293 **Phylogenomic distribution of BclA across the *C. difficile* lineage.**
294 To further explore the diversity of BclA among *C. difficile* lineage, a set of 25,165 *C. difficile* genomes
295 were analyzed for the presence of these proteins (**Table S2**). BclA1 was found in 15,680 (**Table S3**),
296 BclA2 in 21,540 (**Table S4**) and BclA3 in 17,685 genomes (**Table S5**). The distribution of hits in a
297 phylogenetic core-genome tree is shown in Figure 3A. In the case of BclA1 occurrence, a homogeneous

298 distribution is observed in all clades except for clade 4 and clade 5, where BclA1 was not found in most
299 or any of the genomes searched, 74% and 100%, respectively (**Fig. 3A, S14A, S14B & Table S6**). The
300 presence of BclA2 was homogeneous across all clades, the same as BclA3 except for clade 5, where it
301 was not found in 97.6% of the genomes (**Fig. 3A, S14B & Table S6**).
302
303 Clade 1 comprised 16,264 genomes, among which the majority harbored complete open reading frames
304 (ORFs) for *bclA* orthologues: *bclA1* in 68.42%, *bclA2* in 85.75%, and *bclA3* in 74.02% of genomes.
305 Notably, 54.65% (n = 8,889) of genomes in this clade contained all three *bclA* homologs. A smaller subset
306 presented only *bclA2* (10.25%), *bclA3* (3.24%), or *bclA1* (1.53%) individually. Pseudogenized *bclA* genes
307 were rare in this clade, with *bclA1*, *bclA2*, and *bclA3* pseudogenization observed in only 0.1%, 0.2%, and
308 0.17% of genomes, respectively. These results indicate that Clade 1 harbors a largely intact and
309 widespread distribution of *bclA*, with minimal gene inactivation (**Fig 3A, Table S6-S8**).
310
311 Of the 4,565 genomes analyzed of Clade 2, only 56 genomes (1.23%) contained a complete *bclA1* ORF.
312 In contrast, 80.92% (n = 3,694) exhibited a pseudogenized *bclA1*, indicating a clade-wide inactivation of
313 this gene. Despite this, most genomes (69.31%) retained all three *bclA* homologs, including pseudogenized
314 copies. A smaller fraction harbored only one *bclA* ORF (*bclA1*, 2.85%; *bclA2*, 3.5%; *bclA3*, 2.8%), and
315 pseudogenized *bclA2* and bclA3 were found in 0.18% and 0.11% of genomes, respectively. These findings
316 suggest that while *bclA1* is frequently inactivated in Clade 2, the overall *bclA* repertoire remains largely
317 preserved (**Fig 3B, Table S6-S8**).
318
319 Among the 470 genomes analyzed from Clade 3, pseudogenization was widespread, with 81.7% of
320 genomes containing at least one inactivated *bclA* gene. Complete ORFs for *bclA2* and *bclA3* were absent
321 from all genomes, while only three genomes carried a complete *bclA1* (0.6%). Pseudogenized alleles were
322 prevalent: *bclA1* in 84.04%, *bclA2* in 98.72%, and *bclA3* in 93.83% of genomes. Only a small subset
323 contained a single *bclA* gene (*bclA2*, 4.26%, n=20; *bclA3*, 0.21%, n=1), and none harbored *bclA1* alone.
324 This suggest that Clade 3 might not have functional BclA proteins, based on its highly pseudogenized
325 alleles (**Fig 3C, Table S6-S8**).
326
327 Clade 4 included 1,254 genomes. A majority possessed complete ORFs for *bclA2* (90.59%) and *bclA3*
328 (82.62%), while *bclA1* was found intact in only 25.52% of genomes. However, just 17.15% (n = 215) of

329 genomes contained all three *bclA* full length homologs. A minority carried only *bclA1* (0.88%), *bclA2*
330 (6.94%), or *bclA3* (4.55%), and pseudogenization was limited, occurring in 0.32% (*bclA1*), 0% (*bclA2*),
331 and 0.16% (*bclA3*) of genomes. These data show that while *bclA2* and *bclA3* are well-conserved in Clade
332 4, *bclA1* is frequently absent or incomplete, limiting the presence of the full paralog set (**Fig S14A, Table**
333 **S6-S8**).
334
335 Out of the 2,491 genomes of Clade 5, 79.69% possessed a complete *bclA2* ORF. Interestingly, the majority
336 was missing *bclA1* (100%) or bclA3 (97.59%), and a few genomes *bclA2* (19.55%). Pseudogenized alleles
337 were scarce and included *bclA2* in 0.36% and *bclA3* in 0.04%. Overall, these results suggest that Clade 5,
338 with the complete absence of *bclA1* and *bclA3*, will only have BclA2 protein in its repertoire. (**Fig S14B,**
339 **Table S6-S8**).
340
341 Unfortunately, the *C. difficile* data set had few representatives of Clade 6 (n = 5) and Clade 7 (n = 1),
342 therefore it is not possible to determine trends or statistics for these clades (**Table S6**).
343 Regarding the cryptic clades, almost the entirety of C-IIIa and C-IIIb genomes where missing *bclA1* and
344 *bclA3*, and roughly 50% did not presented *bclA2*. Meanwhile, C-II and C-IV genomes did not present
345 *bclA2* at all, contrary to C-V where 43% only presented *bclA2* (**Fig. S15C, Table S6, S7, S9**).
346
347 **Pseudogenization of *bclAs* across *C. difficile* clades.**
348 As stated before, pseudogenization of *bclA1* is present in 80.92% of Clade 2 and 84% of Clade 3 genomes
349 (**Fig. 4A, Table S6-S8**). In the case of *bclA2*, 98.72%, and for *bclA3*, 93.83% of Clade 3 genomes have
350 these genes pseudogenized (**Fig. 4B, 4C & Table S6-S8**). Notably, specific pseudogenized events were
351 highly abundant for each of the *bclAs*. At the sequence level, *bclA1* pseudogenization occurred at residue
352 49 in 99.49% of cases, consistent with the pattern observed in Clade 2 and Clade 3 (**Fig 4D, 4G, Table**
353 **S10**). In the case of *bclA2*, pseudogenization events occurred at residue 86 in 79% of cases, all were
354 present in Clade 3. This *bclA2* allele yielded a truncated protein that retained its NTD and the first 72
355 (17.39%) residues of the CLR (**Fig 4E, 4G, Table S11**). For *bclA3*, pseudogenization events occurred at
356 residue 44, resulting in predicted proteins that retain nearly the full N-terminal domain (NTD, 46 residues),
357 minus two amino acids, all occurring in Clade 3 (**Fig 4F, 4G, Table S12**). These intermediate-length
358 truncations suggest the potential for limited structural function, such as the formation of short, hair-like
359 exosporium projections. Together, these findings indicate that Clade 3 has undergone extensive and

360 consistent *bclA* pseudogenization, likely altering spore surface architecture while preserving partial
361 domain structures in some homologues.
362
363 The presence of low abundant *bclAs* pseudogenes was also studied. These alleles were pseudogenized in
364 the collagen-like region (red lines in CLR region) (**Fig. 4D, 4E & 4F**), and were distributed among the
365 clades. Since these alleles were unique, it is highly likely that these mutations have no selective pressure
366 associated (**Table S10-S12**).
367
368 **Identification and variability of *bclA* alleles**
369 To further identify differences at the amino acid level in the known diversity of *C. difficile* BclAs and
370 determine if there are alleles, a selection of *bclA* genes was used. To do this, deduplicated complete genes
371 were extracted to obtain a set of unique alleles for *bclA1* (1,057, **Table S13**), *bclA2* (1,355, **Table S14**),
372 and *bclA3* (1,842, **Table S15**). For each *bclA* set, sequences were aligned with Clustal Omega against the
373 reference, sequence that presented the least differences compared to all corresponding genes extracted,
374 *bclA1* (CLO_CA1356AA) **(Fig. 5A)**, *bclA2* (CLO_AA3432AA) **(Fig. 5B)** and *bclA3* (CLO_AA0395AA)
375 **(Fig. 5C)**. The alignment was based in codon-aligned amino acids of each BclA set. BclA proteins
376 presented highly conserved NTD and CTD domains with few preserved substitutions events distributed
377 throughout (yellow lines). Remarkably, the CLR region has the most substitutions and Indels occurrence,
378 making it a highly variable region for BclA collagen-like proteins **(Fig. 5).**
379
380 The average amino acid identity was measure for the set of unique BclA alleles. The percentage identity
381 of full length BclA alleles is more than 90% identical in classical clades, but when compared to cryptic
382 clades, a noticeable difference in identity is shown (**Fig S15A, S15B & S15C**). These differences are
383 accentuated when NTD (**Fig. S15D, S15E & S15F**) and CTD (**Fig. S15G, S15H & S15I**) regions are
384 analyzed individually, the identity separation is easily distinguished for all BclA proteins. These findings
385 suggest that the alleles of the classical lineages exhibit a fuzzy separation, and that it is only when
386 compared to the cryptic clades that different types of BclAs become apparent, meaning that classical BclA
387 alleles are very similar. Amino acid identity matrices can be found in Tables S16-S24.

388 **Discussions**

389 *C. difficile* spores possess an outermost exosporium layer that serves as the initial point of contact with
390 the host. Recent research has revealed the ultrastructural diversity, molecular composition, and functional
391 characteristics of this outer layer [11, 13-16, 30, 59-61]. Extensive studies have shown that BclA proteins,
392 distinguished by their collagen-like domains, are key contributors to the assembly and structural
393 organization of the exosporium layer in spores of the *Bacillus cereus* group [31, 34, 62-66]. In contrast, the
394 specific roles and functional implications of BclA-like proteins in *C. difficile* spores remain less clearly
395 defined[14, 16, 17, 52]. Partial characterization in at least two *C. difficile* strains has indicated that these proteins
396 contribute to colonization, interactions with host tissues, spore persistence and recurrence of the disease[14,
397 16, 17]. In this study, we expanded our insights of the BclA in *C. difficile* by performing a comprehensive
398 analysis of the distribution and variability of the *bclA* orthologues family across over 25,000 *C. difficile*
399 genomes from diverse phylogenetic clades. Furthermore, AlphaFold modeling was employed to
400 structurally characterize all three predicted BclA proteins, identifying key conserved and divergent
401 regions. We also addressed frequent pseudogenization events affecting the *bclA* genes, particularly *bclA1*,
402 and experimentally evaluated the functional consequences by reintroducing a full-length *bclA1* gene into
403 a strain with a pseudogenized version. Our findings offer several major conclusions that collectively
404 advance the understanding of the evolutionary dynamics and functional significance of BclA proteins in
405 *C. difficile* spore biology and pathogenesis.

406

407 A major finding of our study is the extensive variability in the presence, sequence conservation, and
408 integrity of *bclA* orthologues across *C. difficile* clades 1 through 5 including the cryptic clades. While
409 *bclA2* and *bclA3* were broadly conserved across most clades, *bclA1* exhibited substantial variability,
410 including complete or partial pseudogenization in multiple lineages—particularly within clade 2 and the
411 cryptic clades. We also observed that clade 5, considered an early ancestor of the four clades, lacked *bclA1*
412 and *bclA3*, while all genomes of clade 3 lack all three *bclA* orthologues. These patterns suggest that
413 selective pressures have driven divergent evolutionary trajectories for spore surface proteins, possibly
414 linked to clade-specific ecological niches or host interactions. It is also tempting to hypothesize that the
415 *bclA* patterns observed across clades might be tied to morphological differences of the hair-like
416 projections. Prior work in strain R20291 that has hair-like projections demonstrates that BclA3 is essential
417 for the formation of the hair-like projections, yet it is unclear whether BclA2 contributes to their
418 formation[16]. Based on these findings, a tentative hypothesis is that all members of clade 3, which lacks all

419 three *bclA* alleles, would be hair-less spores. However, in prior work we demonstrated that spores of clade
420 1 (strain TL178 [ribotype 002] and TL176 [ribotype 014-020]), clade 2 (strain R20291 [RT027]), and
421 clade 5 (strain M120 [ribotype 078-126]) had hair-like projections, irrespective of the *bclA* encoded genes
422 [67]. These observations suggest that the formation of hair-like projections on *C. difficile* spores is more
423 complex than previously understood and may involve differential expression of *bclA* orthologues across
424 clades, potentially compensating for the absence or pseudogenization of one or more genes.
425
426 The high variability observed in *bclA*s across clades also suggests distinct functional redundancy in
427 virulent traits related to spore host interactions. Functional studies in different clades have shed insight
428 into their roles [14, 16, 17]. For example, work in clade 1 strain 630, which lacks hair-like projections[11, 17, 30],
429 shows that *bclA1* is required for proper colonization in mice[17]. However, disruption of individual *bclA*s
430 led to spores lacking the electron-dense exosporium and a defective spore coat, suggesting a pleiotropic
431 effect in the spore surface[17]. By contrast, work on a Clade 2 strain (R20291), shows that a single deletion
432 of *bclA3* is sufficient to impair the formation of the hair-like projection without impacting the formation
433 of the lower electron-dense layers and spore coat[16]. BclA3 was also associated with persistence and disease
434 recurrence[16]. This limited knowledge on *C. difficile* BclAs suggests that depending on the clade, BclAs
435 might have distinct niche-specific advantages, underscoring the need to investigate the impact of BclAs
436 in different clades.
437
438 Another major conclusion offered by this work is the striking pattern of sequence conservation and
439 variability across the BclA proteins-BclA1, BclA2 and BclA3. In each case, both the N-terminal and C-
440 terminal domains were highly conserved across the more than 25,000 *bclA* alleles analyzed, while the
441 central collagen-like regions exhibited substantial sequence diversity. This fact that conserved terminal
442 domains flank a highly variable central region, suggests preservation of the domains requires for BclA
443 anchoring and structural stability (the NTD and CTD, respectively), while allowing evolutionary
444 flexibility in the central collagen-like region. The pronounced variability in the collagen-like domains of
445 all three BclA may be attributed to the repetitive GXY triplet motif that results in underlying DNA
446 sequence repeats. Such repeats are known to increase genetic instability through replication slippage or
447 unequal recombination, which can result in insertions, deletions, or sequence rearrangements[68]. While we
448 did not investigate further in this work, it is likely that in *C. difficile* the collagen-like domain may be
449 subject to diversification to host-derived selective pressures, including evasion host immune responses or

450 adapt to distinct host environments as observed in *B. anthracis* spores[69, 70]. Together, these findings
451 indicate that the collagen-like region is a shared hotspot for adaptive variation across the BclA protein
452 family in *C. difficile*. Additionally, these observations deserve further work to address how variations in
453 the collagen-like region affects spore adherence, environmental persistence, or host colonization
454 dynamics.
455
456 Another notable finding from our analysis is the pattern of *bclA1* pseudogenization across *C. difficile*
457 clades. While pseudogenization events in *bclA* genes were generally rare, occurring at a frequency below
458 1% across most clades, a striking exception was observed in Clade 2. In this clade, most strains carried a
459 conserved pseudogenized *bclA1* allele that encodes only a short 48-amino acid peptide corresponding to
460 the N-terminal domain (NTD) of BclA1. Despite the truncation, this NTD fragment appears to retain the
461 structural features necessary for anchoring to the spore surface[17]. The conservation of this pseudogenized
462 form within Clade 2 is particularly intriguing, as it suggests that full-length BclA1 may be dispensable in
463 this lineage, or that the truncated form fulfills a distinct, potentially clade-specific function. This raises
464 questions about the evolutionary pressures shaping *bclA1* function and whether the presence of only the
465 anchoring domain represents an adaptive modification.
466
467 In summary, this work highlights the diversity of the collagen-like BclA proteins in *C. difficile* and
468 demonstrate that they exhibit conserved modular domains defined by a highly conserved N- and C-
469 terminal domains, flanked by a highly variable collagen-like region. While the precise implications of
470 Clade-specific variabilities remains unclear,
471

472 **Figure legend**

473

474 **Fig. 1. BclA domain structure in *C. difficile* 630 & R20291 strain and AlphaFold3 BclA CTD**
475 **predictions. (A)** Schematic representation of *C. difficile* 630 and R20291 BclA1, BclA2 and BclA3
476 protein sequences. NTD (Black): N-terminal domain, CLR (Maroon): Collagen-like region, CTD
477 (Yellow): C-terminal domain. In red arrows are depicted two stop codons that led to early termination of
478 translation of $BclA1_{R20291}$. **(B, D, F)** AlphaFold3 prediction of CTDs monomers of $BclA1_{630}$, $BclA2_{R20291}$
479 and $BclA3_{R20291}$, and topology models to the right are shown. α-helix and loops are displayed in brown
480 and black, respectively. Yellow β-sheets are located on the inner face of homotrimers models. Grey β-
481 strands are located on the outer face of homotrimers models. Encircled "N" and "C" represent the amino-
482 and carboxyl-terminus, respectively. Structural predictions indicate that $BclA1_{630}$ had a total of 12 β-
483 sheets, $BclA2_{R20291}$ had a total of 9 β-sheets and $BclA3_{R20291}$ encodes 14 β-sheets and 1 α-helix. **(C, E, G)**
484 AlphaFold3 prediction of CTDs homotrimers of $BclA1_{630}$, $BclA2_{R20291}$ and $BclA3_{R20291}$ and **(H)** their
485 Predicted aligned error (PAE) are shown.

486

487 **Fig. 2. *C. difficile* BclA comparison with BclA and BclB CTD domain from *B. anthracis*.** Alphafold3
488 predicted CTD homotrimers structures of *C. difficile* (Cd) and *B. anthracis* (Ba) were superimposed on
489 ChimeraX (v.1.9) using "matchmaker" command with default parameters. **(A)** RMSD for both pruned
490 and all atom pairs is reported for all combinations possible. Numbers in bold represent the lowest RMSD
491 achieved for each protein comparison. Superimposed pairs are **(B)** $Cd_BclA1_CTD_{630}$ with
492 Ba_BclA_CTD, **(C)** $Cd_BclA2_CTD_{R20291}$ with Ba_BclA_CTD and **(D)** $Cd_BclA3_CTD_{R20291}$ with
493 Ba_BclB_CTD. *C. difficile* 630 genome used as reference was GenBank ID: AM180355.1. *C. difficile*
494 R20291 genome used as reference was GenBank ID: CP029423. *B. anthracis* genome used as reference
495 was GenBank ID: AE016879.1. RMSD: Root mean square deviation, Cd: *Clostridioides difficile*, Ba:
496 *Bacillus anthracis.*

497

498 **Fig 3. Distribution of BclAs in known diversity of *C. difficile*. (A)** Phylogenetic tree of 25k core-genome
499 of *C. difficile*, showing the clade each genome belongs to. Horizontal bars represent the state of BclA in a
500 particular genome (not found, complete, containing Ns, interrupted, or pseudogenized). **(B)** Zoomed-in
501 view of the Clade 2 subtree. **(C)** Zoomed-in view of the Clade 3 subtree.

502

503 **Fig. 4. State of BclAs in known diversity of *C. difficile*.** Prevalence of found conditions of **(A)** BclA1,
504 **(B)** BclA2 and **(C)** BclA3 on each clade. Prevalence of pseudogenized **(D)** BclA1, **(E)** BclA2 and **(F)**
505 BclA3 sequences per each *C. difficile* clade. Schematic representation of BclAs protein sequences are
506 depicted on top of each donut chart. Truncated BclA sequences were manually curated to locate position
507 of early stop codons, represented as red lines in a white rectangle below protein schematic. NTD (Black):
508 N-terminal domain, CLR (Maroon): Collagen-like region, CTD (Yellow): C-terminal domain. In this sub-
509 data set, only clades represented in the figure are present in the legend.
510
511 **Fig. 5. Comparison of *C. difficile* BclA1, BclA2, and BclA3 unique alleles.** Unique alleles were aligned
512 using Clustal Omega against the reference. Truncated, interrupted or pseudogenized sequences were not
513 considered in this analysis. Heatmap displaying positions that differ or not from the reference sequence
514 are depicted for **(A)** BclA1, Ref genome: CLO_CA1356AA, **(B)** BclA2, Ref genome: CLO_AA3432AA
515 and **(C)** BclA3, Ref genome: CLO_AA0395AA. Dashed red lines mark the division between BclA
516 domains. N-terminal domain (NTD, black), Collagen-like region (CLR, maroon) and C-terminal domain
517 (CTD, yellow).
518
519
520
521
522

523 **Supplementary figure legend**

524

525 **Fig. S1. *C. difficile bclA1* gene schematic representation in 630 & R20291 strain.** *bclA1*$_{630}$
526 (CD630_03320) is a 2082 bp gene coding a 694 amino acid protein (67.8 kDa). *bclA1*$_{R20291}$
527 (CDIF27147_00471, CDIF27147_00472, CDIF27147_00473) is represented as a three-segmented Open
528 Reading Frame (ORF) (*bclA1_1, bclA1_2 & bclA1_3*) because it is a pseudogenized gene. A nonsense
529 mutation, A145T, generated an early stop codon resulting in a short protein of 48 amino acids (4.7 kDa).
530 Downstream, another nonsense mutation, C739T, also generates an early stop codon and the two extra
531 segments of *bclA1*$_{R20291}$ ORF. *bclA1* possess a σ^K promoter region (Saujet, 2013), and two putative σ^E and
532 σ^G consensus sequences that were found during this work. *bclA1* is flanked upstream by *ppiB* and
533 downstream by *ppaC*. *C. difficile* 630 genome used as reference was GenBank ID: AM180355.1. *C.*
534 *difficile* R20291 genome used as reference was GenBank ID: CP029423. Scale bar: 1 kb.

535

536 **Fig. S2. BclA1 Pairwise sequence alignment (PSA) of 630 & R20291 strains.** EMBOSS Needle was
537 used for PSA. A representation of the protein domains is depicted above the sequence. The proteins
538 possess a 51.4% Identity. Repaired BclA1 from R20291 was used for the alignment. N-terminal domain
539 (NTD, Black), Collagen-like region (CLR, Maroon) and C-terminal domain (CTD, Yellow). *C. difficile*
540 630 genome used as reference was GenBank ID: AM180355.1. *C. difficile* R20291 genome used as
541 reference was GenBank ID: CP029423.

542

543 **Fig. S3. *C. difficile bclA2* gene schematic representation in 630 & R20291 strains.** *bclA2*$_{630}$
544 (CD630_32300) is a 1677 bp gene coding a 558 amino acid protein (49.1 kDa). *bclA2*$_{R20291}$
545 (CDIF27147_03409) is a 1641 bp gene coding a 546 amino acid protein (47.9 kDa). *bclA2* possess a σ^K
546 promoter region (Saujet, 2013) and a putative σ^E consensus sequence that was found during this work.
547 *bclA2* is flanked upstream by *hpt2* and downstream by *dapB2*. *C. difficile* 630 genome used as reference
548 was GenBank ID: AM180355.1. *C. difficile* R20291 genome used as reference was GenBank ID:
549 CP029423. Scale bar: 1 kb.

550

551 **Fig. S4. BclA2 Pairwise sequence alignment (PSA) of 630 & R20291 strains.** EMBOSS Needle was
552 used for PSA. A representation of the protein domains is depicted above the sequence. The proteins
553 possess a 95.2% Identity. N-terminal domain (NTD, Black), Collagen-like region (CLR, Maroon) and C-

554 terminal domain (CTD, Yellow). *C. difficile* 630 genome used as reference was GenBank ID:
555 AM180355.1. *C. difficile* R20291 genome used as reference was GenBank ID: CP029423.
556
557 **Fig. S5. *C. difficile bclA3* gene schematic representation in 630 & R20291 strains.** *bclA3*$_{630}$
558 (CD630_33490) is a 1986 bp gene coding a 661 amino acid protein (58.3 kDa). *bclA3*$_{R20291}$
559 (CDIF27147_03519) is a 2037 bp gene coding a 678 amino acid protein (59.9 kDa). *bclA3* is the second
560 gene of an operon with an upstream *sgtA* gene coding for a glycosyl transferase (CD630_33500,
561 CDIF27147_03520). The operon possesses a σ^K promoter region, and an additional σ^K promoter region
562 (Saujet, 2013) immediately upstream *bclA3* and a putative σ^E consensus sequence that was found upstream
563 the operon during this work. *bclA3* is flanked upstream by *sgtA* and downstream by CDIF27147_03518.
564 *C. difficile* 630 genome used as reference was GenBank ID: AM180355.1. *C. difficile* R20291 genome
565 used as reference was GenBank ID: CP029423. Scale bar: 1 kb.
566
567 **Fig. S6. BclA3 Pairwise sequence alignment (PSA) of 630 & R20291 strains.** EMBOSS Needle was
568 used for PSA. A representation of the protein domains is depicted above the sequence. The proteins
569 possess an 86.9% Identity. N-terminal domain (NTD, Black), Collagen-like region (CLR, Maroon) and
570 C-terminal domain (CTD, Yellow). *C. difficile* 630 genome used as reference was GenBank ID:
571 AM180355.1. *C. difficile* R20291 genome used as reference was GenBank ID: CP029423.
572
573 **Fig. S7. Full length BclA1 Alphafold3 predicted structure.** Predicted monomeric structure of *C. difficile*
574 **(A)** BclA1$_{630}$, **(B)** rBclA1$_{R20291}$, and predicted homotrimer structures of **(C)** BclA1$_{630}$ and **(D)**
575 rBclA1$_{R20291}$. Each monomer unit has a different color for easy recognition in trimeric structures. Each
576 model has its respective Predicted aligned error (PAE) plot below. NTD: N-terminal domain, CLR:
577 Collagen-like region, CTD: C-terminal domain.
578
579 **Fig. S8. Full length BclA2$_{R20291}$ Alphafold3 predicted structure.** Predicted structure of *C. difficile* **(A)**
580 BclA2$_{R20291}$ monomer and **(B)** homotrimer complex. Each monomer unit has a different color for easy
581 recognition in trimeric structures. Each model has its respective Predicted aligned error (PAE) plot below.
582 NTD: N-terminal domain, CLR: Collagen-like region, CTD: C-terminal domain.
583

584 **Fig. S9. Full length BclA3$_{R20291}$ Alphafold3 predicted structure.** Predicted structure of *C. difficile* **(A)**
585 BclA3$_{R20291}$ monomer and **(B)** homotrimer complex. Each monomer unit has a different color for easy
586 recognition in trimeric structures. Each model has its respective Predicted aligned error (PAE) plot below.
587 NTD: N-terminal domain, CLR: Collagen-like region, CTD: C-terminal domain.
588
589 **Fig. S10. AlphaFold3 predicted structure of BclA_10GXY homotrimer.** Predicted structure of *C.*
590 *difficile* **(A)** BclA1$_{630}$, **(B)** BclA2$_{R20291}$ and **(C)** BclA3$_{R20291}$ homotrimer complexes. Each model has its
591 respective Predicted aligned error (PAE) plot below.
592
593 **Fig. S11. AlphaFold3 predicted CTD heterotrimer structures from R20291 strain.** Heterotrimer
594 predicted model composed of **(A)** 2 units of BclA2_CTD and 1 unit of BclA3_CTD or **(B)** 1 unit of
595 BclA2_CTD and 2 unit of BclA3_CTD. Each model has its respective Predicted aligned error (PAE) plot
596 below.
597
598 **Fig. S12. AlphaFold3 predicted CTD heterotrimer structures from 630 strain.** Heterotrimer
599 prediction composed of **(A)** 2 units of BclA1_CTD and 1 unit of BclA2_CTD, **(B)** 2 units of BclA1_CTD
600 and 1 unit of BclA3_CTD, **(C)** 1 unit of BclA1_CTD and 2 unit of BclA2_CTD, **(D)** 1 unit of BclA1_CTD
601 and 2 unit of BclA3_CTD or **(E)** 1 unit of BclA1_CTD, 1 unit of BclA2_CTD and 1 unit of BclA3_CTD.
602 Each model has its respective Predicted aligned error (PAE) plot.
603
604 **Fig. S13. Pairwise alignment (PSA) of *C. difficile* BclA_CTD with *B. anthracis* BclA/BclB_CTD.**
605 EMBOSS Needle was used for PSA of **(A)** Cd_BclA1_CTD$_{630}$ with Ba_BclA_CTD **(B)**
606 Cd_BclA2_CTD$_{R20291}$ with Ba_BclA_CTD and **(C)** Cd_BclA3_CTD$_{R20291}$ with Ba_BclB_CTD. **(D)**
607 Percentage of identity of aligned sequences. CTD: C-terminal domain. *C. difficile* 630 genome used as
608 reference was GenBank sequence: AM180355.1. *C. difficile* R20291 genome used as reference was
609 GenBank sequence: CP029423. *B. anthracis* genome used as reference was GenBank sequence:
610 AE016879.1.
611
612 **Fig. S14. Distribution of BclAs in known diversity of *C. difficile*. (A)** Zoomed-in view of the Clade 4
613 subtree. **(B)** Zoomed-in view of the Clade 5 subtree. **(C)** Zoomed-in view of the cryptic clades subtree.

614 Horizontal bars represent the state of BclA in a particular genome (not found, complete, containing Ns,
615 interrupted, or pseudogenized).
616
617 **Fig. S15. Average amino acid identity (AAI) analysis of BclA unique alleles.** Heatmap of the
618 percentage of identity between all unique alleles of **(A)** BclA1, **(B)** BclA2 and **(C)** BclA3. AAI was
619 performed for full length protein, NTD and CTD regions. N-terminal domain (NTD) and C-terminal
620 domain (CTD).
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637

638 **References**

639

640

641

642 1. Rupnik M, Wilcox MH, Gerding DN. Clostridium difficile infection: new developments in
643 epidemiology and pathogenesis. Nat Rev Microbiol. 2009;7(7):526-36. doi:
644 10.1038/nrmicro2164. PubMed PMID: 19528959.
645 2. Martin JS, Monaghan TM, Wilcox MH. Clostridium difficile infection: epidemiology,
646 diagnosis and understanding transmission. Nat Rev Gastroenterol Hepatol. 2016;13(4):206-16.
647 Epub 20160309. doi: 10.1038/nrgastro.2016.25. PubMed PMID: 26956066.
648 3. Dingle KE, Elliott B, Robinson E, Griffiths D, Eyre DW, Stoesser N, Vaughan A, Golubchik
649 T, Fawley WN, Wilcox MH, Peto TE, Walker AS, Riley TV, Crook DW, Didelot X. Evolutionary
650 history of the Clostridium difficile pathogenicity locus. Genome Biol Evol. 2014;6(1):36-52. doi:
651 10.1093/gbe/evt204. PubMed PMID: 24336451; PMCID: PMC3914685.
652 4. Finn E, Andersson FL, Madin-Warburton M. Burden of Clostridioides difficile infection
653 (CDI) - a systematic review of the epidemiology of primary and recurrent CDI. BMC Infect Dis.
654 2021;21(1):456. Epub 20210519. doi: 10.1186/s12879-021-06147-y. PubMed PMID: 34016040;
655 PMCID: PMC8135979.
656 5. Lessa FC, Mu Y, Bamberg WM, Beldavs ZG, Dumyati GK, Dunn JR, Farley MM,
657 Holzbauer SM, Meek JI, Phipps EC, Wilson LE, Winston LG, Cohen JA, Limbago BM, Fridkin
658 SK, Gerding DN, McDonald LC. Burden of Clostridium difficile infection in the United States. N
659 Engl J Med. 2015;372(9):825-34. doi: 10.1056/NEJMoa1408913. PubMed PMID: 25714160.
660 6. Marsh JW, Arora R, Schlackman JL, Shutt KA, Curry SR, Harrison LH. Association of
661 relapse of Clostridium difficile disease with BI/NAP1/027. J Clin Microbiol. 2012;50(12):4078-82.
662 Epub 20121010. doi: 10.1128/JCM.02291-12. PubMed PMID: 23052318; PMCID:
663 PMC3502988.
664 7. Schaffler H, Breitruck A. Clostridium difficile - From Colonization to Infection. Front
665 Microbiol. 2018;9:646. Epub 20180410. doi: 10.3389/fmicb.2018.00646. PubMed PMID:
666 29692762; PMCID: PMC5902504.
667 8. Aktories K, Schwan C, Jank T. *Clostridium difficile* toxin biology. Annu Rev Microbiol.
668 2017;71:281-307. Epub 2017/06/29. doi: 10.1146/annurev-micro-090816-093458. PubMed
669 PMID: 28657883.
670 9. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. Clostridium difficile infection. Nat
671 Rev Dis Primers. 2016;2:16020. Epub 20160407. doi: 10.1038/nrdp.2016.20. PubMed PMID:
672 27158839; PMCID: PMC5453186.
673 10. Deakin LJ, Clare S, Fagan RP, Dawson LF, Pickard DJ, West MR, Wren BW, Fairweather
674 NF, Dougan G, Lawley TD. The *Clostridium difficile spo0A* gene is a persistence and
675 transmission factor. Infect Immun. 2012;80(8):2704-11. Epub 2012/05/23. doi:
676 10.1128/IAI.00147-12. PubMed PMID: 22615253; PMCID: PMC3434595.
677 11. Paredes-Sabja D, Shen A, Sorg JA. Clostridium difficile spore biology: sporulation,
678 germination, and spore structural proteins. Trends Microbiol. 2014;22(7):406-16. Epub
679 20140507. doi: 10.1016/j.tim.2014.04.003. PubMed PMID: 24814671; PMCID: PMC4098856.

680 12. Paredes-Sabja D, Sarker MR. Adherence of *Clostridium difficile* spores to Caco-2 cells in
681 culture. J Med Microbiol. 2012;61(Pt 9):1208-18. Epub 2012/05/19. doi: 10.1099/jmm.0.043687-
682 0. PubMed PMID: 22595914.
683 13. Mora-Uribe P, Miranda-Cardenas C, Castro-Cordova P, Gil F, Calderon I, Fuentes JA,
684 Rodas PI, Banawas S, Sarker MR, Paredes-Sabja D. Characterization of the Adherence of
685 *Clostridium difficile* Spores: The Integrity of the Outermost Layer Affects Adherence Properties
686 of Spores of the Epidemic Strain R20291 to Components of the Intestinal Mucosa. Front Cell
687 Infect Microbiol. 2016;6:99. Epub 2016/10/08. doi: 10.3389/fcimb.2016.00099. PubMed PMID:
688 27713865; PMCID: PMC5031699.
689 14. Castro-Córdova P, Otto-Medina M, Montes-Bravo N, Brito-Silva C, Lacy DB, Paredes-
690 Sabja D. Redistribution of the Novel Clostridioides difficile Spore Adherence Receptor E-
691 Cadherin by TcdA and TcdB Increases Spore Binding to Adherens Junctions. Infect Immun.
692 2023;91(1):e0047622. Epub 20221130. doi: 10.1128/iai.00476-22. PubMed PMID: 36448839;
693 PMCID: PMC9872679.
694 15. Calderon-Romero P, Castro-Cordova P, Reyes-Ramirez R, Milano-Cespedes M,
695 Guerrero-Araya E, Pizarro-Guajardo M, Olguin-Araneda V, Gil F, Paredes-Sabja D. *Clostridium*
696 *difficile* exosporium cysteine-rich proteins are essential for the morphogenesis of the exosporium
697 layer, spore resistance, and affect *C. difficile* pathogenesis. PLoS Pathog. 2018;14(8):e1007199.
698 Epub 2018/08/09. doi: 10.1371/journal.ppat.1007199. PubMed PMID: 30089172.
699 16. Castro-Córdova P, Mora-Uribe P, Reyes-Ramírez R, Cofré-Araneda G, Orozco-Aguilar
700 J, Brito-Silva C, Mendoza-León MJ, Kuehne SA, Minton NP, Pizarro-Guajardo M, Paredes-Sabja
701 D. Entry of spores into intestinal epithelial cells contributes to recurrence of *Clostridioides difficile*
702 infection. Nat Commun. 2021;12(1):1140. Epub 2021/02/18. doi: 10.1038/s41467-021-21355-5.
703 PubMed PMID: 33602902.
704 17. Phetcharaburanin J, Hong HA, Colenutt C, Bianconi I, Sempere L, Permpoonpattana P,
705 Smith K, Dembek M, Tan S, Brisson MC, Brisson AR, Fairweather NF, Cutting SM. The spore-
706 associated protein BclA1 affects the susceptibility of animals to colonization and infection by
707 *Clostridium difficile*. Mol Microbiol. 2014;92(5):1025-38. Epub 2014/04/12. doi:
708 10.1111/mmi.12611. PubMed PMID: 24720767.
709 18. Knight DR, Imwattana K, Kullin B, Guerrero-Araya E, Paredes-Sabja D, Didelot X, Dingle
710 KE, Eyre DW, Rodríguez C, Riley TV. Major genetic discontinuity and novel toxigenic species
711 in *Clostridioides difficile* taxonomy. Elife. 2021;10. Epub 2021/06/11. doi: 10.7554/eLife.64325.
712 PubMed PMID: 34114561; PMCID: PMC8241443.
713 19. Knight DR, Elliott B, Chang BJ, Perkins TT, Riley TV. Diversity and Evolution in the
714 Genome of Clostridium difficile. Clin Microbiol Rev. 2015;28(3):721-41. doi: 10.1128/cmr.00127-
715 14. PubMed PMID: 26085550; PMCID: PMC4475645.
716 20. Knight DR, Kullin B, Androga GO, Barbut F, Eckert C, Johnson S, Spigaglia P, Tateda K,
717 Tsai PJ, Riley TV. Evolutionary and Genomic Insights into Clostridioides difficile Sequence Type
718 11: a Diverse Zoonotic and Antimicrobial-Resistant Lineage of Global One Health Importance.
719 mBio. 2019;10(2). Epub 20190416. doi: 10.1128/mBio.00446-19. PubMed PMID: 30992351;
720 PMCID: PMC6469969.
721 21. Mengoli M, Barone M, Fabbrini M, D'Amico F, Brigidi P, Turroni S. Make It Less difficile:
722 Understanding Genetic Evolution and Global Spread of Clostridioides difficile. Genes (Basel).
723 2022;13(12). Epub 20221124. doi: 10.3390/genes13122200. PubMed PMID: 36553467;
724 PMCID: PMC9778335.

725 22. Clements AC, Magalhães RJ, Tatem AJ, Paterson DL, Riley TV. Clostridium difficile PCR
726 ribotype 027: assessing the risks of further worldwide spread. Lancet Infect Dis. 2010;10(6):395-
727 404. doi: 10.1016/s1473-3099(10)70080-3. PubMed PMID: 20510280; PMCID: PMC7185771.
728 23. Endres BT, Begum K, Sun H, Walk ST, Memariani A, Lancaster C, Gonzales-Luna AJ,
729 Dotson KM, Bassères E, Offiong C, Tupy S, Kuper K, Septimus E, Arafat R, Alam MJ, Zhao Z,
730 Hurdle JG, Savidge TC, Garey KW. Epidemic Clostridioides difficile Ribotype 027 Lineages:
731 Comparisons of Texas Versus Worldwide Strains. Open Forum Infect Dis. 2019;6(2):ofz013.
732 Epub 20190209. doi: 10.1093/ofid/ofz013. PubMed PMID: 30793006; PMCID: PMC6368847.
733 24. McDonald LC, Killgore GE, Thompson A, Owens RC, Jr., Kazakova SV, Sambol SP,
734 Johnson S, Gerding DN. An epidemic, toxin gene-variant strain of Clostridium difficile. N Engl J
735 Med. 2005;353(23):2433-41. Epub 20051201. doi: 10.1056/NEJMoa051590. PubMed PMID:
736 16322603.
737 25. Li C, Harmanus C, Zhu D, Meng X, Wang S, Duan J, Liu S, Fu C, Zhou P, Liu R, Wu A,
738 Kuijper EJ, Smits WK, Fu L, Sun X. Characterization of the virulence of a non-RT027, non-
739 RT078 and binary toxin-positive Clostridium difficile strain associated with severe diarrhea.
740 Emerg Microbes Infect. 2018;7(1):211. Epub 20181212. doi: 10.1038/s41426-018-0211-1.
741 PubMed PMID: 30542069; PMCID: PMC6291415.
742 26. Imwattana K, Rodríguez C, Riley TV, Knight DR. A species-wide genetic atlas of
743 antimicrobial resistance in Clostridioides difficile. Microb Genom. 2021;7(11). doi:
744 10.1099/mgen.0.000696. PubMed PMID: 34793295; PMCID: PMC8743556.
745 27. Kumar N, Browne HP, Viciani E, Forster SC, Clare S, Harcourt K, Stares MD, Dougan G,
746 Fairley DJ, Roberts P, Pirmohamed M, Clokie MRJ, Jensen MBF, Hargreaves KR, Ip M, Wieler
747 LH, Seyboldt C, Norén T, Riley TV, Kuijper EJ, Wren BW, Lawley TD. Adaptation of host
748 transmission cycle during Clostridium difficile speciation. Nat Genet. 2019;51(9):1315-20. Epub
749 20190812. doi: 10.1038/s41588-019-0478-8. PubMed PMID: 31406348.
750 28. Janezic S, Potocnik M, Zidaric V, Rupnik M. Highly Divergent Clostridium difficile Strains
751 Isolated from the Environment. PLoS One. 2016;11(11):e0167101. Epub 20161123. doi:
752 10.1371/journal.pone.0167101. PubMed PMID: 27880843; PMCID: PMC5120845.
753 29. Williamson CHD, Stone NE, Nunnally AE, Roe CC, Vazquez AJ, Lucero SA, Hornstra H,
754 Wagner DM, Keim P, Rupnik M, Janezic S, Sahl JW. Identification of novel, cryptic Clostridioides
755 species isolates from environmental samples collected from diverse geographical locations.
756 Microb Genom. 2022;8(2). doi: 10.1099/mgen.0.000742. PubMed PMID: 35166655; PMCID:
757 PMC8942030.
758 30. Paredes-Sabja D, Cid-Rojas F, Pizarro-Guajardo M. Assembly of the exosporium layer in
759 Clostridioides difficile spores. Curr Opin Microbiol. 2022;67:102137. Epub 20220216. doi:
760 10.1016/j.mib.2022.01.008. PubMed PMID: 35182899.
761 31. Steichen C, Chen P, Kearney JF, Turnbough CL. Identification of the immunodominant
762 protein and other proteins of the *Bacillus anthracis* exosporium. J Bacteriol. 2003;185(6):1903-
763 10. doi: 10.1128/jb.185.6.1903-1910.2003. PubMed PMID: 12618454; PMCID: PMC150129.
764 32. Sylvestre P, Couture-Tosi E, Mock M. A collagen-like surface glycoprotein is a structural
765 component of the Bacillus anthracis exosporium. Mol Microbiol. 2002;45(1):169-78. doi:
766 10.1046/j.1365-2958.2000.03000.x. PubMed PMID: 12100557.
767 33. Strong PC, Fulton KM, Aubry A, Foote S, Twine SM, Logan SM. Identification and
768 characterization of glycoproteins on the spore surface of Clostridium difficile. J Bacteriol.
769 2014;196(14):2627-37. Epub 20140509. doi: 10.1128/jb.01469-14. PubMed PMID: 24816601;
770 PMCID: PMC4097583.

34. Boydston JA, Chen P, Steichen CT, Turnbough CL. Orientation within the exosporium and structural stability of the collagen-like glycoprotein BclA of *Bacillus anthracis*. J Bacteriol. 2005;187(15):5310-7. doi: 10.1128/JB.187.15.5310-5317.2005. PubMed PMID: 16030225; PMCID: PMC1196033.
35. Maes E, Krzewinski F, Garenaux E, Lequette Y, Coddeville B, Trivelli X, Ronse A, Faille C, Guerardel Y. Glycosylation of BclA Glycoprotein from Bacillus cereus and Bacillus anthracis Exosporium Is Domain-specific. J Biol Chem. 2016;291(18):9666-77. Epub 20160226. doi: 10.1074/jbc.M116.718171. PubMed PMID: 26921321; PMCID: PMC4850304.
36. Réty S, Salamitou S, Garcia-Verdugo I, Hulmes DJ, Le Hégarat F, Chaby R, Lewit-Bentley A. The crystal structure of the Bacillus anthracis spore surface protein BclA shows remarkable similarity to mammalian proteins. J Biol Chem. 2005;280(52):43073-8. Epub 20051025. doi: 10.1074/jbc.M510087200. PubMed PMID: 16249180.
37. Lukomski S, Nakashima K, Abdi I, Cipriano VJ, Ireland RM, Reid SD, Adams GG, Musser JM. Identification and characterization of the scl gene encoding a group A Streptococcus extracellular protein virulence factor with similarity to human collagen. Infect Immun. 2000;68(12):6542-53. doi: 10.1128/iai.68.12.6542-6553.2000. PubMed PMID: 11083763; PMCID: PMC97748.
38. Squeglia F, Bachert B, De Simone A, Lukomski S, Berisio R. The crystal structure of the streptococcal collagen-like protein 2 globular domain from invasive M3-type group A Streptococcus shows significant similarity to immunomodulatory HIV protein gp41. J Biol Chem. 2014;289(8):5122-33. Epub 20131219. doi: 10.1074/jbc.M113.523597. PubMed PMID: 24356966; PMCID: PMC3931070.
39. Xu Y, Keene DR, Bujnicki JM, Höök M, Lukomski S. Streptococcal Scl1 and Scl2 proteins form collagen-like triple helices. J Biol Chem. 2002;277(30):27312-8. Epub 20020425. doi: 10.1074/jbc.M201163200. PubMed PMID: 11976327.
40. Brodsky B, Persikov AV. Molecular structure of the collagen triple helix. Adv Protein Chem. 2005;70:301-39. doi: 10.1016/s0065-3233(05)70009-7. PubMed PMID: 15837519.
41. Knight DR, Imwattana K, Kullin B, Guerrero-Araya E, Paredes-Sabja D, Didelot X, Dingle KE, Eyre DW, Rodriguez C, Riley TV. Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. Elife. 2021;10. Epub 20210611. doi: 10.7554/eLife.64325. PubMed PMID: 34114561; PMCID: PMC8241443.
42. Guerrero-Araya E, Munoz M, Rodriguez C, Paredes-Sabja D. FastMLST: A Multi-core Tool for Multilocus Sequence Typing of Draft Genome Assemblies. Bioinform Biol Insights. 2021;15:11779322211059238. Epub 20211127. doi: 10.1177/11779322211059238. PubMed PMID: 34866905; PMCID: PMC8637782.
43. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124. Epub 20180924. doi: 10.12688/wellcomeopenres.14826.1. PubMed PMID: 30345391; PMCID: PMC6192448.
44. Buchfink B, Reuter K, Drost HG. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nat Methods. 2021;18(4):366-8. Epub 20210407. doi: 10.1038/s41592-021-01101-x. PubMed PMID: 33828273; PMCID: PMC8026399.
45. Hyatt D, Chen GL, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics. 2010;11:119. Epub 20100308. doi: 10.1186/1471-2105-11-119. PubMed PMID: 20211023; PMCID: PMC2848648.

46. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30(4):772-80. Epub 20130116. doi: 10.1093/molbev/mst010. PubMed PMID: 23329690; PMCID: PMC3603318.
47. Didelot X, Wilson DJ. ClonalFrameML: efficient inference of recombination in whole bacterial genomes. PLoS Comput Biol. 2015;11(2):e1004041. Epub 20150212. doi: 10.1371/journal.pcbi.1004041. PubMed PMID: 25675341; PMCID: PMC4326465.
48. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, Lanfear R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37(5):1530-4. doi: 10.1093/molbev/msaa015. PubMed PMID: 32011700; PMCID: PMC7182206.
49. Boudko SP, Engel J, Bächinger HP. The crucial role of trimerization domains in collagen folding. Int J Biochem Cell Biol. 2012;44(1):21-32. Epub 20111005. doi: 10.1016/j.biocel.2011.09.009. PubMed PMID: 22001560.
50. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Soding J, Thompson JD, Higgins DG. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539. Epub 20111011. doi: 10.1038/msb.2011.75. PubMed PMID: 21988835; PMCID: PMC3261699.
51. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22(13):1658-9. Epub 20060526. doi: 10.1093/bioinformatics/btl158. PubMed PMID: 16731699.
52. Pizarro-Guajardo M, Olguin-Araneda V, Barra-Carrasco J, Brito-Silva C, Sarker MR, Paredes-Sabja D. Characterization of the collagen-like exosporium protein, BclA1, of *Clostridium difficile* spores. Anaerobe. 2014;25:18-30. Epub 2013/11/26. doi: 10.1016/j.anaerobe.2013.11.003. PubMed PMID: 24269655.
53. Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, Bodenstein SW, Evans DA, Hung CC, O'Neill M, Reiman D, Tunyasuvunakool K, Wu Z, Žemgulytė A, Arvaniti E, Beattie C, Bertolli O, Bridgland A, Cherepanov A, Congreve M, Cowen-Rivers AI, Cowie A, Figurnov M, Fuchs FB, Gladman H, Jain R, Khan YA, Low CMR, Perlin K, Potapenko A, Savy P, Singh S, Stecula A, Thillaisundaram A, Tong C, Yakneen S, Zhong ED, Zielinski M, Žídek A, Bapst V, Kohli P, Jaderberg M, Hassabis D, Jumper JM. Accurate structure prediction of biomolecular interactions with AlphaFold 3. Nature. 2024;630(8016):493-500. Epub 20240508. doi: 10.1038/s41586-024-07487-w. PubMed PMID: 38718835; PMCID: PMC11168924.
54. Wirz JA, Boudko SP, Lerch TF, Chapman MS, Bächinger HP. Crystal structure of the human collagen XV trimerization domain: a potent trimerizing unit common to multiplexin collagens. Matrix Biol. 2011;30(1):9-15. Epub 20101013. doi: 10.1016/j.matbio.2010.09.005. PubMed PMID: 20932905; PMCID: PMC3048825.
55. Xu Y, Kirchner M. Collagen Mimetic Peptides. Bioengineering (Basel). 2021;8(1). Epub 20210105. doi: 10.3390/bioengineering8010005. PubMed PMID: 33466358; PMCID: PMC7824840.
56. Sakakibara S, Inouye K, Shudo K, Kishida Y, Kobayashi Y, Prockop DJ. Synthesis of (Pro-Hyp-Gly) n of defined molecular weights. Evidence for the stabilization of collagen triple helix by hydroxypyroline. Biochim Biophys Acta. 1973;303(1):198-202. doi: 10.1016/0005-2795(73)90164-5. PubMed PMID: 4702003.

861 57. Ricard-Blum S. The collagen family. Cold Spring Harb Perspect Biol. 2011;3(1):a004978.
862 Epub 20110101. doi: 10.1101/cshperspect.a004978. PubMed PMID: 21421911; PMCID:
863 PMC3003457.
864 58. Sharma S, Bhatnagar R, Gaur D. Complement Evasion Strategies of Human Pathogenic
865 Bacteria. Indian J Microbiol. 2020;60(3):283-96. Epub 20200424. doi: 10.1007/s12088-020-
866 00872-9. PubMed PMID: 32655196; PMCID: PMC7329968.
867 59. Pizarro-Guajardo M, Ravanal MC, Paez MD, Callegari E, Paredes-Sabja D. Identification
868 of *Clostridium difficile* Immunoreactive Spore Proteins of the Epidemic Strain R20291.
869 Proteomics Clin Appl. 2018:e1700182. Epub 2018/03/25. doi: 10.1002/prca.201700182.
870 PubMed PMID: 29573213.
871 60. Díaz-González F, Milano M, Olguin-Araneda V, Pizarro-Cerda J, Castro-Córdova P,
872 Tzeng SC, Maier CS, Sarker MR, Paredes-Sabja D. Protein composition of the outermost
873 exosporium-like layer of Clostridium difficile 630 spores. J Proteomics. 2015;123:1-13. Epub
874 20150404. doi: 10.1016/j.jprot.2015.03.035. PubMed PMID: 25849250; PMCID: PMC6764588.
875 61. Pizarro-Guajardo M, Olguín-Araneda V, Barra-Carrasco J, Brito-Silva C, Sarker MR,
876 Paredes-Sabja D. Characterization of the collagen-like exosporium protein, BclA1, of Clostridium
877 difficile spores. Anaerobe. 2014;25:18-30. Epub 20131121. doi:
878 10.1016/j.anaerobe.2013.11.003. PubMed PMID: 24269655.
879 62. Thompson BM, Hsieh HY, Spreng KA, Stewart GC. The co-dependence of BxpB/ExsFA
880 and BclA for proper incorporation into the exosporium of *Bacillus anthracis*. Mol Microbiol.
881 2011;79(3):799-813. doi: 10.1111/j.1365-2958.2010.07488.x. PubMed PMID: 21255119;
882 PMCID: PMC3044595.
883 63. Thompson BM, Stewart GC. Targeting of the BclA and BclB proteins to the *Bacillus*
884 *anthracis* spore surface. Mol Microbiol. 2008;70(2):421-34. Epub 2008/08/27. doi:
885 10.1111/j.1365-2958.2008.06420.x. PubMed PMID: 18761690.
886 64. Thompson BM, Waller LN, Fox KF, Fox A, Stewart GC. The BclB glycoprotein of *Bacillus*
887 *anthracis* is involved in exosporium integrity. J Bacteriol. 2007;189(18):6704-13. Epub
888 2007/07/20. doi: 10.1128/JB.00762-07. PubMed PMID: 17644587; PMCID: PMC2045162.
889 65. Steichen CT, Kearney JF, Turnbough CL. Characterization of the exosporium basal layer
890 protein BxpB of *Bacillus anthracis*. J Bacteriol. 2005;187(17):5868-76. doi:
891 10.1128/JB.187.17.5868-5876.2005. PubMed PMID: 16109927; PMCID: PMC1196169.
892 66. Sylvestre P, Couture-Tosi E, Mock M. Polymorphism in the collagen-like region of the
893 *Bacillus anthracis* BclA protein leads to variation in exosporium filament length. J Bacteriol.
894 2003;185(5):1555-63. doi: 10.1128/jb.185.5.1555-1563.2003. PubMed PMID: 12591872;
895 PMCID: PMC148075.
896 67. Pizarro-Guajardo M, Calderon-Romero P, Castro-Cordova P, Mora-Uribe P, Paredes-
897 Sabja D. Ultrastructural Variability of the Exosporium Layer of *Clostridium difficile* Spores. Appl
898 Environ Microbiol. 2016;82(7):2202-9. Epub 2016/02/07. doi: 10.1128/AEM.03410-15. PubMed
899 PMID: 26850296; PMCID: PMC4807528.
900 68. Brazda V, Fojta M, Bowater RP. Structures and stability of simple DNA repeats from
901 bacteria. Biochem J. 2020;477(2):325-39. doi: 10.1042/bcj20190703. PubMed PMID: 31967649;
902 PMCID: PMC7015867.
903 69. Wang Y, Jenkins SA, Gu C, Shree A, Martinez-Moczygemba M, Herold J, Botto M, Wetsel
904 RA, Xu Y. Bacillus anthracis Spore Surface Protein BclA Mediates Complement Factor H
905 Binding to Spores and Promotes Spore Persistence. PLoS Pathog. 2016;12(6):e1005678. Epub

906 20160615. doi: 10.1371/journal.ppat.1005678. PubMed PMID: 27304426; PMCID:
907 PMC4909234.
908 70. Xue Q, Gu C, Rivera J, Hook M, Chen X, Pozzi A, Xu Y. Entry of *Bacillus anthracis* spores
909 into epithelial cells is mediated by the spore surface protein BclA, integrin alpha2beta1 and
910 complement component C1q. Cell Microbiol. 2011;13(4):620-34. Epub 2010/12/08. doi:
911 10.1111/j.1462-5822.2010.01558.x. PubMed PMID: 21134100.
912

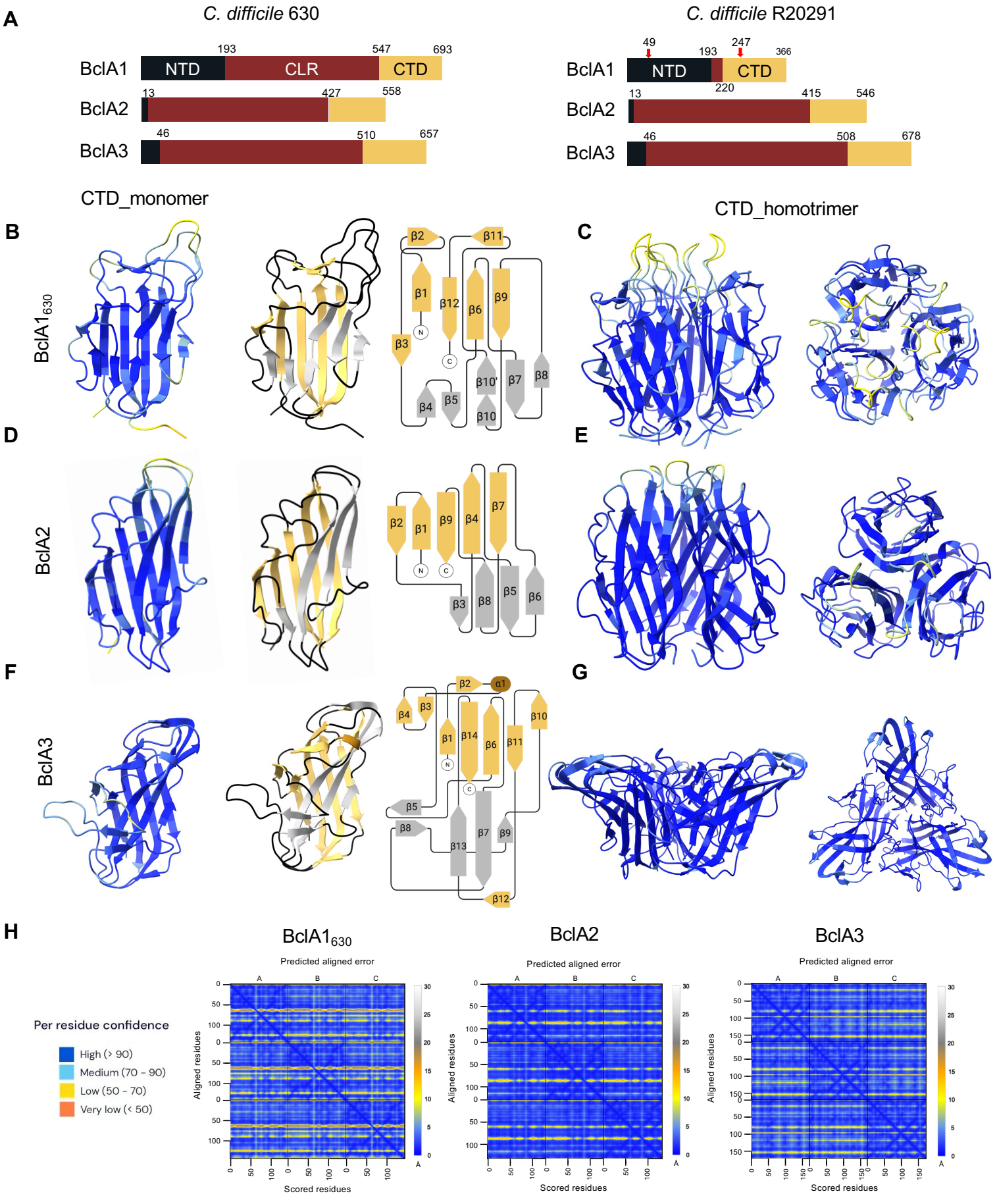
A
C. difficile 630
C. difficile R20291
BclA1
NTD
CLR
CTD
193
547
693
49
247
366
220
BclA2
13
427
558
415
546
BclA3
46
510
657
508
678
CTD_monomer
CTD_homotrimer
B
C
D
E
F
G
BclA1630
BclA2
BclA3
H
Predicted aligned error
Aligned residues
Scored residues
Per residue confidence
High (> 90)
Medium (70 – 90)
Low (50 – 70)
Very low (< 50)

Fig. 1. BclA domain structure in *C. difficile* 630 & R20291 strain and AlphaFold3 BclA CTD predictions. (A) Schematic representation of *C. difficile* 630 and R20291 BclA1, BclA2 and BclA3 protein sequences. NTD (Black): N-terminal domain, CLR (Maroon): Collagen-like region, CTD (Yellow): C-terminal domain. In red arrows are depicted two stop codons that led to early termination of translation of $BclA1_{R20291}$. **(B,D,F)** AlphaFold3 prediction of CTDs monomers of $BclA1_{630}$, $BclA2_{R20291}$ and $BclA3_{R20291}$, and topology models to the right are shown. α-helix and loops are displayed in brown and black, respectively. Yellow β-sheets are located on the inner face of homotrimers models. Grey β-strands are located on the outer face of homotrimers models. Encircled "N" and "C" represent the amino- and carboxyl-terminus, respectively. Structural predictions indicate that $BclA1_{630}$ had a total of 12 β-sheets, $BclA2_{R20291}$ had a total of 9 β-sheets and $BclA3_{R20291}$ encodes 14 β-sheets and 1 α-helix. **(C,E,G)** AlphaFold3 prediction of CTDs homotrimers of $BclA1_{630}$, $BclA2_{R20291}$ and $BclA3_{R20291}$ and **(H)** their Predicted aligned error (PAE) are shown.

A

C. difficile	*B. anthracis* BclA_CTD	*B. anthracis* BclB_CTD
Pruned atom pairs (RMSD)		
BclA1_CTD	**60 (0.956)**	8 (0.928)
BclA2_CTD	**46 (1.078)**	15 (1.382)
BclA3_CTD	10 (1.187)	**120 (0.745)**
All atom pairs (RMSD)		
BclA1_CTD	122 (3.940)	118 (16.692)
BclA2_CTD	113 (6.480)	117 (12.652)
BclA3_CTD	110 (17.972)	158 (2.808)

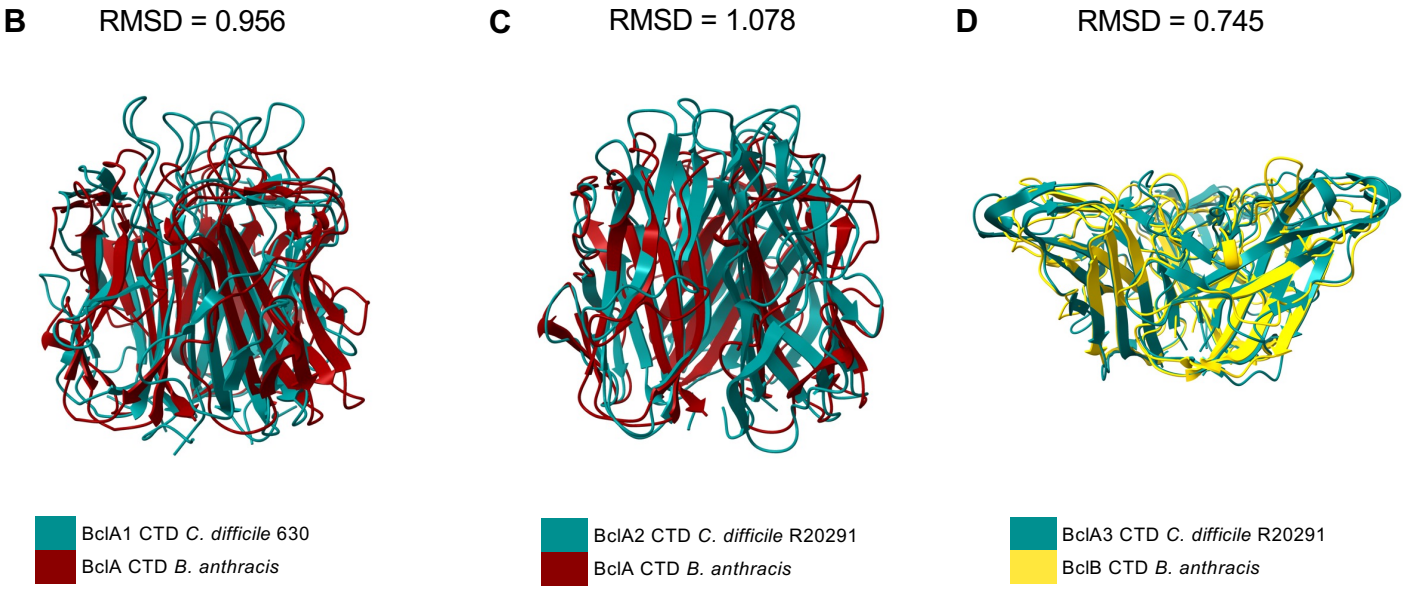

Fig. 2. *C. difficile* BclA comparison with BclA and BclB CTD domain from *B. anthracis*. Alphafold3 predicted CTD homotrimers structures of *C. difficile* (Cd) and *B. anthracis* (Ba) were superimposed on ChimeraX (v.1.9) using "matchmaker" command with default parameters. **(A)** RMSD for both pruned and all atom pairs is reported for all combinations possible. Numbers in bold represent the lowest RMSD achieved for each protein comparison. Superimposed pairs are **(B)** Cd_BclA1_CTD$_{630}$ with Ba_BclA_CTD, **(C)** Cd_BclA2_CTD$_{R20291}$ with Ba_BclA_CTD and **(D)** Cd_BclA3_CTD$_{R20291}$ with Ba_BclB_CTD. *C. difficile* 630 genome used as reference was GenBank ID: AM180355.1. *C. difficile* R20291 genome used as reference was GenBank ID: CP029423. *B. anthracis* genome used as reference was GenBank ID: AE016879.1. RMSD: Root mean square deviation, Cd: *Clostridioides difficile*, Ba: *Bacillus anthracis.*

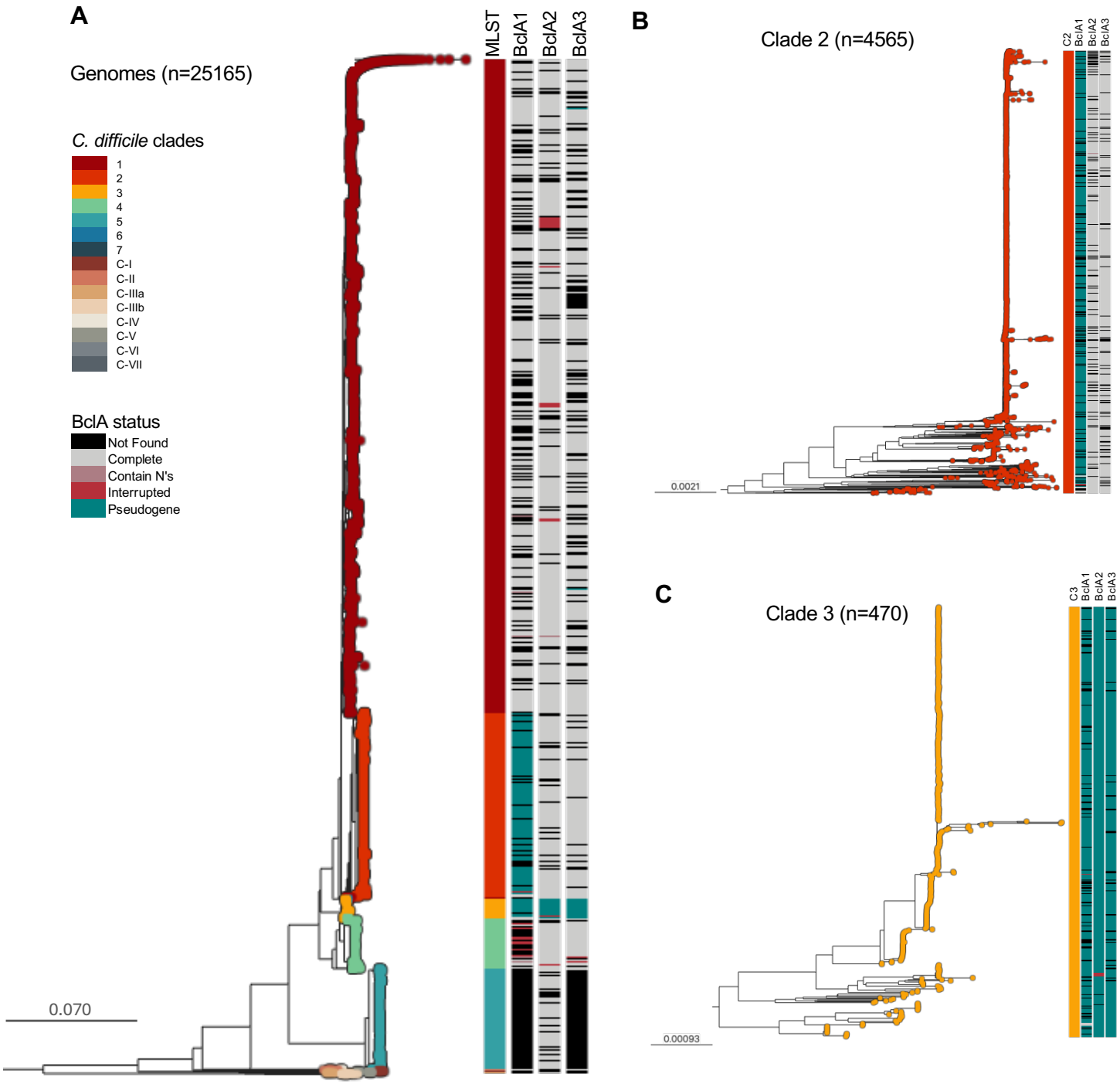

Fig 3. Distribution of BclAs in known diversity of *C. difficile*. (A) Phylogenetic tree of 25k core-genome of *C. difficile*, showing the clade each genome belongs to. Horizontal bars represent the state of BclA in a particular genome (not found, complete, containing Ns, interrupted, or pseudogenized). **(B)** Zoomed-in view of the Clade 2 subtree. **(C)** Zoomed-in view of the Clade 3 subtree

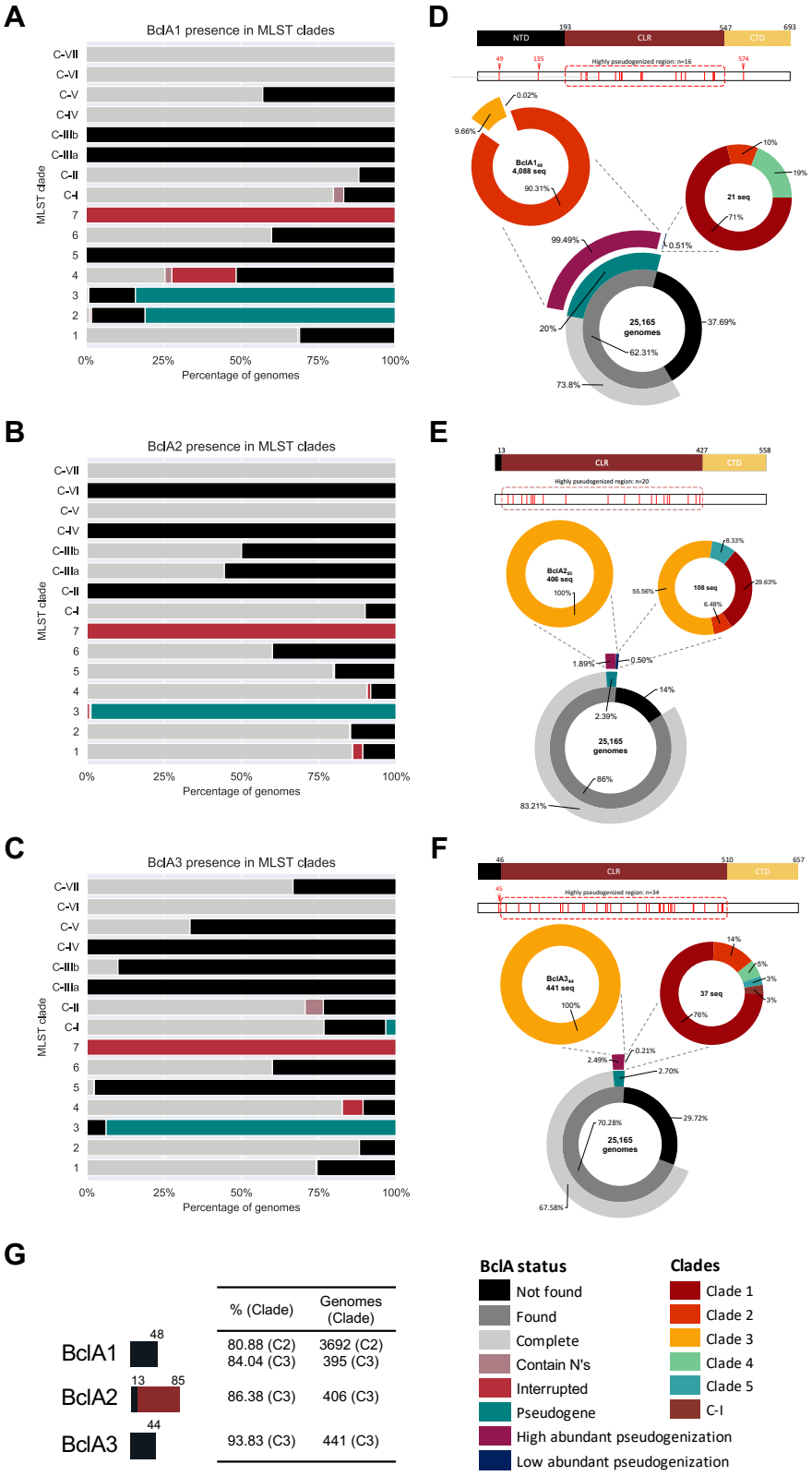

Fig. 4. State of BclAs in known diversity of *C. difficile*. Prevalence of found conditions of **(A)** BclA1, **(B)** BclA2 and **(C)** BclA3 on each clade. Prevalence of pseudogenized **(D)** BclA1, **(E)** BclA2 and **(F)** BclA3 sequences per each *C. difficile* clade. Schematic representation of BclAs protein sequences are depicted on top of each donut chart. Truncated BclA sequences were manually curated to locate position of early stop codons, represented as red lines in a white rectangle below protein schematic. **(G)** BclA high abundant pseudogenization alleles present in clade 2 and clade 3. NTD (Black): N-terminal domain, CLR (Maroon): Collagen-like region, CTD (Yellow): C-terminal domain. In this sub-data set, only clades represented in the figure are present in the legend.

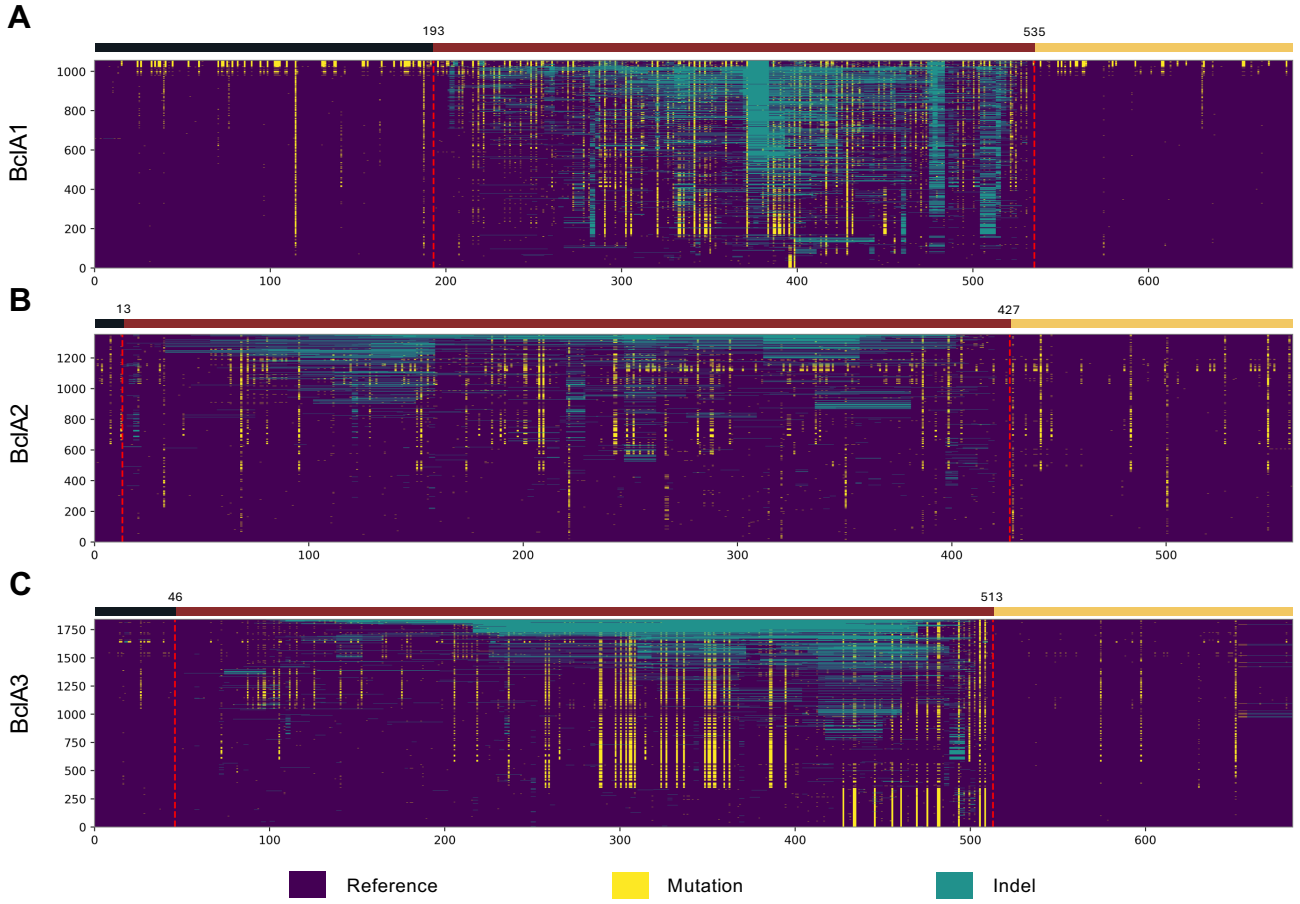

Fig. 5. Comparison of *C. difficile* BclA1, BclA2, and BclA3 unique alleles. Unique alleles were aligned using Clustal Omega against the reference. Truncated, interrupted or pseudogenized sequences were not considerer in this analysis. Heatmap displaying positions that differ or not from the reference sequence are depicted for **(A)** BclA1, Ref genome: CLO_CA1356AA, **(B)** BclA2, Ref genome: CLO_AA3432AA and **(C)** BclA3, Ref genome: CLO_AA0395AA. Dashed red lines mark the division between BclA domains. N-terminal domain (NTD, black), Collagen-like region (CLR, maroon) and C-terminal domain (CTD, yellow).