1 ***Clostridioides difficile*** **PCR ribotype 151 is polyphyletic and includes pathogenic isolates from**

2 **cryptic clade C-II with mono-toxin pathogenicity loci that can escape routine diagnostics**

3

4 Quinten R. Ducarmon[a,b], Tjomme van der Bruggen[c], Céline Harmanus[a], Ingrid M.J.G. Sanders[a], Laura

5 G.M. Daenen[d], Ad C. Fluit[c], Rolf H.A.M. Vossen[e], Susan L. Kloet[e], Ed J. Kuijper[a,b,f], Wiep Klaas

6 Smits[a,b,g#]

7

8 [a]Department of Medical Microbiology, Leiden University Medical Center, Leiden, The Netherlands.

9 [b]Center for Microbiome Analyses and Therapeutics, Leiden University Medical Center, Leiden, The

10 Netherlands.

11 [c]Department of Medical Microbiology, University Medical Centre Utrecht, Utrecht, The Netherlands.

12 [d]Department of Haematology, University Medical Centre Utrecht, Utrecht, The Netherlands.

13 [e]Leiden Genome Technology Center, Leiden University Medical Center, Leiden, The Netherlands

14 [f]Center for Infectious Disease Control, National Institute for Public Health and the Environment,

15 Bilthoven, The Netherlands.

16 [g]Centre for Microbial Cell Biology, Leiden, The Netherlands.

17 [#]Address correspondence to w.k.smits@lumc.nl

18

ABSTRACT

We report a patient case with pseudomembranous colitis associated with a mono-toxin producing *Clostridioides difficile* belonging to the very rarely diagnosed PCR ribotype (RT) 151. The infection was difficult to diagnose, since the isolate and the feces sample tested negative for toxin-encoding genes using a routine commercial test. This prompted us to sequence n = 11 RT151s from various geographical regions to study their genomic characteristics and relatedness. By including whole genome sequence data from other sources, we could further place these isolates into the phylogenetic tree of *C. difficile* and assign them to their respective clades. These analyses revealed that 1) RT151s are polyphyletic with isolates falling into clades 1, and cryptic clades C- I and C- II 2) RT151 contains both non-toxigenic and toxigenic isolates and 3) RT151 C-II isolates contained mono-toxin pathogenicity loci (PaLoc). Additional analysis with PacBio circular consensus sequencing revealed that the isolate from our patient case report contains a novel PaLoc insertion site, lacked *tcdA* and a had significantly divergent *tcdB* sequence that might explain the failure of the diagnostic test. The study is noteworthy as 1) RT151 encompasses both typical and cryptic clades and 2) conclusive evidence for CDI due to clade C-II isolates was hitherto lacking.

KEYWORDS: *Clostridioides difficile*, cryptic clade, PaLoc

36 **INTRODUCTION**

37 *Clostridioides difficile* is a Gram-positive, spore-forming anaerobic bacterium and the leading cause
38 of antibiotic-associated diarrhea, both in healthcare facilities and the community(1). *C. difficile* can
39 produce several clostridial toxins (toxin A or TcdA, toxin B or TcdB and the binary toxin CDT) and
40 these are responsible for *C. difficile* infection (CDI), symptoms of which can range from mild and self-
41 limiting diarrhea to pseudomembranous colitis, toxic megacolon and ultimately death(1).
42 Phylogenetic analyses have revealed that most *C. difficile* isolates associated with human and animal
43 disease fall within 5 clades(2). In recent years, reports have appeared on (mainly environmental)
44 isolates that are phylogenetically distinct from these 5 clades, but rather fall within at least 3 so-
45 called cryptic clades (C-I, C-II and C-III) that can be considered a separate genomospecies(2-5).
46 Though cryptic clade isolates can contain toxin genes(2, 4, 6), knowledge about the occurrence and
47 clinical significance of these isolates is limited.

48 Within clades 1-5, the pathogenicity locus harboring the the *tcdA* and/or *tcdB* genes is
49 generally located at a consistent location in the *C. difficile* genome (*cdu1*/*cdd1*) and further includes
50 several other accessory genes (*tcdR*, *tcdC* and *tcdE*) (6, 7). Importantly, isolates belonging to the
51 cryptic clades C-I and C-III can have divergent pathogenicity locus (PaLoc) sequences and atypical
52 PaLoc insertion sites that can impact diagnostic tests(2, 4, 6, 8).

53 Here, we report a case of human CDI caused by a RT151 clade C-II isolate for which initial
54 diagnostic tests using Cepheid Xpert *C. difficile* BT assay were negative, even though *C. difficile* could
55 be cultured from patient feces and clinical symptoms corresponded to a CDI case. We show that the
56 cryptic C-II isolate cultured from this patient contains a mono-toxin B PaLoc at a novel location, and
57 that other RT151 isolates fall in phylogenetically divergent clades 1, C-I and C-II.

58

59 **MATERIALS AND METHODS**

60 ***Clostridioides difficile* culturing, ribotyping and toxin testing**

61 RT151 isolates analysed as part of this study were derived from various national and international
62 studies in which the Dutch National Expertise Center for *C. difficile* at Leiden University Medical
63 Center (LUMC) participated and had isolates available. The isolates originate from Greece, Malta,
64 The Netherlands and Spain between 2011 and 2021. Detailed information on all RT151 isolates,
65 including relevant accession numbers, sequence types (STs), clade and metadata is provided in
66 Supplemental Table S1.

67 Isolates were cultured on TSS plates (Tryptic Soy Agar with 5% sheep blood, bioMérieux, The
68 Netherlands) or CLO plates (selective *C. difficile* medium containing cefoxitin, amphotericin B and
69 cycloserine, bioMérieux, The Netherlands). Capillary electrophoresis PCR ribotyping was performed
70 at the Dutch Expertise Centre for *C. difficile*, according to standard procedures(9). Feces samples and
71 *C. difficile* isolates were analyzed by the Xpert *C. difficile* BT assay (Cepheid), that targets *tcdB*, *cdtA*
72 and a variant *tcdC* gene associated with epidemic strains, according to the manufacturer's
73 instructions. They were further characterized using an in-house multiplex PCR targeting the 16S
74 rRNA gene, *gluD*, and the genes encoding the large clostrididal toxins and binary toxin(10) or a PCR
75 targeting *tcdB* alone(11). Toxin status was additionally assessed using a toxin neutralization assay
76 using 0.4 µM filter-sterilized culture supernatant from a 48-h culture in Brain-Heart Infusion (BHI)
77 broth on VeroE6 cells, using purified toxins (SML1153-2UG, Lot No SLBT4085, Sigma) and
78 neutralizing anti-TcdA/B antitoxin (T1000, Lot No 1015235, Techlab) (12). Supernatants from 48-h
79 BHI cultures were also analyzed using the VIDAS® *C. difficile* Toxin A & B (Biomérieux), according to
80 the manufacturer's instructions.

81

82 **DNA extraction and whole-genome sequencing**

83 A pure culture from a TSS plate was resuspended in phosphate-buffered saline (Media Products
84 Groningen) to 3 McFarland, and 24 µl of lysozyme (50 g L^{-1}) was added to 800 µl of this suspension.
85 The solution was incubated for 20 min at 37°C and 15 µl of proteinase K (20 g L^{-1}l; Roche) was added,
86 followed by a 20 min incubation at 37°C. DNA was extracted on the QiaSymphony platform (Qiagen,
87 The Netherlands), using the Qiasymphony Virus/Pathogen Midi Kit according to the manufacturer's
88 instructions. Isolated total DNA was submitted for sequencing to Genomescan (Leiden, The
89 Netherlands) on the Illumina Novaseq6000 platform with read length 150 bp in paired-end mode.

90 To generate complete genomes, three isolates (LUMCMM21 0001/0003, LUMCMM19 2333
91 and LUMCMM16 0013) were submitted for long read sequencing on the Pacific Biosciences (PacBio)
92 Sequel platform at the Leiden Genome Technology Center. To prepare high molecular weight total
93 DNA, isolates were cultured overnight in 10 mL BHI at 37°C. Cells from 5 mL of culture were pelleted
94 and processed using the Qiagen Genomic-tip 100/G, according to the manufacturer's instructions.
95 PacBio sequencing libraries were generated as follows. Genomic DNA was sheared with the
96 Megaruptor 3 system (Diagenode) using 35-40 cycles. SMRTbell® libraries were generated according
97 to the following manufacturer's procedure & checklist: Preparing whole genome and metagenome
98 libraries using SMRTbell® prep kit 3.0, PN 102-166-600 APR2022. To retain DNA corresponding to
99 plasmids <10 kb, no size-selection was applied. The libraries were sequenced on a PacBio Sequel II
100 platform with a 30 h movie time using Binding Kit v2.2, Sequencing Primer v5 and Sequencing Kit
101 v2.0.

102

103 ***C. difficile* genome retrieval**

104 Four additional RT151 genome sequences were obtained from the laboratory collection of the UK
105 Anaerobic Reference Unit (Supplemental Table S1); these were processed and sequenced previously
106 as part of their routine surveillance and typing activities but raw Illumina reads were processed in

107 parallel with the other RT151 genomes reported here for the purpose of this study, as described
108 below.

109 To assess the phylogenetic relationship of the eleven RT151 in the context of the previously
110 established clades of *C. difficile*, we downloaded additional isolate sequence data: we included two
111 isolates of each of the three cryptic clades (C-I, C-II and C-III)(2), 4 isolates of clade 1 (as this is the
112 largest group) and two isolates of clade 2-5 from the Leiden-Leeds collection(13, 14) (Supplemental
113 Table S1).

114

115 **Data processing, assembly and quality control**

116 Raw Illumina sequence data was cleaned and trimmed using fastp (v0.23.2)(15) and sequence
117 quality was inspected using Fastqc (v0.11.9) and Multiqc (v1.8)(16, 17). Cleaned reads were
118 assembled by using SKESA (v2.4.0) and SPAdes (v3.15.3) with "–untrusted-contigs" and "–isolate"
119 parameters(18, 19). Only contigs with length 500 >= bp were kept. Assembly quality and general
120 characteristics of the assemblies were assessed using QUAST (v5.0.2)(20). The scaffolds obtained
121 from SPAdes were used for downstream analyses. Default parameters were used unless stated
122 otherwise.

123 Raw PacBio sequence data from LUMCMM16 0013 (RT151, clade C-I), LUMCMM21 0001
124 (RT151, clade C-II) and LUMCMM19 2333 (RT151, clade C-II) was assembled using Flye (v2.9)(21)"
125 and the start position was fixed using Circlator and the "fixstart" parameter(22). Assembly quality
126 was subsequently inspected using available output metrics from Flye, QUAST and completeness
127 using BUSCO (v5.3.2, dataset clostridia_odb10 creation date 2020-03-06)(23). Analysis of m4C and
128 m6A methylation was performed using PacBio SMRT® Link Software (v11.0.0.146107).

129

130 ***C. difficile* genome annotations and sequence typing**

131 Prokka (v1.14.6) was used for rapid genome annotation and to obtain protein sequences, with the "–
132 kingdom Bacteria", "--genus Clostridioides" and "–species difficile" parameters(24). For pan-genome
133 and core-genome analyses, Roary (v3.13.0) was ran on the PROKKA-annotated scaffolds with the "-e
134 mafft" parameters(25, 26).

135 The accessory genome was subsequently visualized in R (v4.2.0) using the ComplexHeatmap
136 package (v2.12.0)(27). To construct a phylogenetic tree based on the core genes as identified by
137 Roary, we used IQTree (v2.2.0) with the "-B 2500" and "-m MFP" parameters for 2500 bootstrap
138 replicates and automated model selection(28, 29). The best model was chosen based on the
139 minimized BIC score and was identified to be UNREST+F0+I+G4. The obtained phylogenetic tree was
140 subsequently visualized using iTOL v6 and the tree was rooted at midpoint(30).

141 Multi-locus sequence typing (MLST) was performed using mlst (v2.19.0) with the PubMLST *C.*
142 *difficile* database updated to October 21st 2021(31). To assign clades to each of our sequence types,
143 we used the information contained in Knight et al, which includes a clade assignment for sequence
144 types(2). Novel loci (sodA-116, recA-82 and glyA-140) were identified and their respective sequences
145 obtained by mlst using the "-q –novel parameters" and submitted to the pubmlst database. Based
146 on this, three isolates were assigned new sequence types (ST946, ST947 and ST948) through
147 PubMLST (https://pubmlst.org/). Lastly, pairwise average nucleotide identities (ANIs) between the
148 genomes were calculated using the fastANI tool (v1.33)(32).

149 Available metadata for all isolates, including newly assigned sequence types, are listed in
150 Supplemental Table 1.

151

152

153 **Data availability**

154 Raw sequence data is available at the European Nucleotide Archive under BioProject PRJEB52887.
155 Associated metadata can be found in Supplemental Table S1. All bioinformatic tools used for the
156 analyses are freely available through the references provided. Complete chromosomes for PacBio-
157 sequenced isolates can be found as accession numbers GCA_945861085 (LUMCMM21 0003),
158 GCA_945909465 (LUMCMM19 2333) and GCA_945909635 (LUMCMM16 0013). For PaLoc alignment
159 and insertion analyses, the corresponding regions were extracted from the complete genome
160 sequence of the reference strain *C. difficile* 630 (AM180355)(33). The pHSJD-312 sequence was
161 obtained from accession MG973074(4). The patient provided written informed consent for the
162 anonymized materials and results to be used for publication.

163

164 **RESULTS AND DISCUSSION**

165 **Case description and identification of a RT151 isolate from a CDI patient**

166 In 2021, a 46-year-old female patient was diagnosed with acute myeloid leukemia and treated with
167 two courses of induction chemotherapy followed by allogeneic hematopoietic stem cell
168 transplantation (allo-HSCT) at the University Medical Center Utrecht (UMCU), a tertiary care center
169 in the Netherlands. Due to primary graft failure, the patient received a second allo-HSCT two months
170 later. Between the first allo-HSCT and engraftment after the second allo-HSCT, the patient was
171 neutropenic for a prolonged time. This period was complicated by candidemia with *Candida kefyr*,
172 treated with anidulafungin followed by fluconazole, two bacteraemia episodes with a multidrug
173 resistant *Klebsiella pneumoniae* treated with imipenem/cilistatin and probable pulmonary
174 aspergillosis for which the patient received antifungal treatment. The post-transplantation period
175 was complicated by early reactivation of cytomegalovirus treated with foscarnet and later letermovir
176 prophylaxis, and mild graft-versus-host disease of the skin for which the patient received topical
177 steroids.

178 In the week prior to the second allo-HSCT, the patient developed diarrhea (day 0) and feces
179 was positive for norovirus and sapovirus as tested by PCR. The commonly used molecular test for
180 *Clostridioides difficile* (Cepheid Xpert *C. difficile* BT) was negative on a stool sample at day 0. Despite
181 the use of loperamide, diarrhea persisted in the following weeks although subsequent PCRs for
182 noro- and sapovirus became negative. PCRs for rotavirus, enterovirus, *Salmonella*, *Shigella*, *Yersinia*,
183 *Campylobacter*, *Plesiomonas*, *Giardia, Cryptosporidium*, *Dientamoeba fragilis*, *Blastocystis hominis*,
184 *Entamoeba histolytica* were also negative. The Xpert *C. difficile* BT assay was repeatedly negative on
185 day 43, day 45 and day 49. Colon biopsies taken on day 53 showed no signs of graft-versus-host
186 disease or pseudomembranous colitis. Around day 50 to 55, diarrhea seemed to diminish, but
187 around day 65 it started to intensify with up to ten watery stools per day in the ensuing weeks.
188 Repeated stool tests for viral, bacterial and parasitic infection were negative, including microscopy

189 for helminth eggs, cysts and *Cyclospora* and the Xpert *C. difficile* BT (on day 66, day 80, day 84 and
190 day 86). However, macroscopic findings during colonoscopy performed on day 82 were suggestive of
191 pseudomembranous colitis, confirmed by microscopy of colon biopsies while there were no signs of
192 graft-versus-host disease. These findings prompted treatment with oral vancomycin starting on day
193 84 using a taper-and-pulse regimen. After 4-5 days, the patient started to improve considerably with
194 stools becoming less frequent, lower stool volumes and a more solid consistency. Ethanol-treated
195 fecal material obtained on day 84 was culture-positive for *C. difficile* on cycloserine cefetoxin
196 fructose agar after 48h anaerobic incubation, as confirmed by matrix-assisted laser
197 desorption/ionization time-of-flight mass spectrometry (MALDI-TOF score of 2.09, database BDAL-
198 8468, Bruker Daltonics, Germany). On day 159 (35 days after completion of anti-CDI treatment), the
199 patient developed a mild episode of diarrhea with abdominal cramps at home, with negative Xpert
200 *C. difficile* BT test on stool but again a positive culture of *C. difficile* from feces (MALDI-TOF score of
201 2.05) that was later confirmed to belong to RT151. The patient spontaneously recovered without
202 antibiotic treatment within the two weeks that followed. Both isolates tested negative in the Xpert
203 *C. difficile* BT assay, but positive when using an in–house developed PCR for *C. difficile* toxin B(11).
204 Both isolates were then sent to the Dutch Expertise Centre for *C. difficile* at the LUMC, with the first
205 isolate receiving the identification number LUMCMM21 0001 and the second isolate LUMCMM21
206 0003. Both isolates were characterized by capillary PCR ribotyping at the LUMC and were shown to
207 be RT151.
208
209 **RT151 isolates have different toxin profiles**

210 To obtain a better insight into the prevalence and characteristics of RT151, we retrieved all isolates
211 from the collection of the Dutch Expertise Center for *C. difficile*, that encompasses > 22000 isolates
212 collected since 2005. Seven isolates had been typed as RT151 using either agarose-gel based
213 ribotyping or capillary electrophoresis-based PCR ribotyping, demonstrating that this is a very rarely
214 diagnosed PCR ribotype (<0.035%).

215 Within this collection of RT151 isolates, five were non-toxigenic *C. difficile* (NTCD) and two
216 were toxigenic on the basis of a multiplex PCR. The isolate obtained from the case description from
217 Utrecht (LUMCMM21 0001) was positive for the gene encoding toxin B, but not the gene encoding
218 toxin A (A^-B^+), while an isolate obtained from a patient in Malta in 2019 (LUMCMM19 2333) had a
219 mono-toxin A (A^+B^-) PaLoc and neither isolate contained the binary toxin locus according to PCR
220 results.

221 These results are in line with findings of the Leeds Teaching Hospitals NHS Trust and
222 University of Leeds that report a prevalence for RT151 of <0.03% with varying toxin status in a cell
223 cytotoxicity and neutralization assay (CCNA; personal communication K.A. Davies and J. Freeman). In
224 addition, the global *C. difficile* analysis carried out by Knight and colleagues also found a prevalence
225 of 0.033% for ST205, based on a collection of 12,098 genomes(2).

226 We also assessed toxigenicity using a diagnostic enzyme-linked fluorescent assay (VIDAS®,
227 bioMérieux) and a VeroE6-based CCNA. The result showed that only the *tcdB*-positive LUMCMM21
228 0001/0003 tested positive for toxin production (Supplemental Table S1). No toxin production could
229 be demonstrated for the *tcdA*-containing isolate (LUMCMM19 2333) under the conditions tested.

230

231 **Core genome phylogeny places some RT151 isolates in cryptic clade C-I and C-II**

232 To determine the relatedness between our RT151 isolates and to understand their placement in the
233 phylogenetic tree of *C. difficile*, we performed Illumina short-read whole genome sequencing (WGS).
234 In addition to the seven RT151 isolates obtained from our in-house collection, the UK Anaerobic
235 Reference Unit provided WGS data from four additional RT151 isolates. From a detailed analysis it
236 became clear that the two isolates isolated from the patient described above (LUMCMM21 0001 and
237 LUMCMM21 0003) are identical (data not shown).

238 By placing the eleven RT151 isolates into a phylogenetic tree with representative isolates
239 from cryptic clades I - III clades and all 'classical' clades to ensure genetic diversity in our tree, we
240 assigned the RT151 isolates to their respective clades based on 1158 core genes in our analysis
241 (Figure 1). This analysis revealed that the majority (n = 8) of our RT151 isolates belonged to Clade 1
242 and MLST typing revealed these to be ST205 (or highly related ST type(s); ST205-like) (Supplemental
243 Table 1 and Figure 1). All of the RT151 isolates obtained from the UK Anaerobic Reference Unit fell
244 into ST205 (or ST205-like) groups and were therefore not further analyzed. Interestingly, three of the
245 isolates were classified into cryptic clades, with two independent isolates belonging to clade C-II
246 (LUMCMM21 0001/0003 and LUMCMM19 2333) and one isolate to clade C-I (LUMCMM16 0013). To
247 complement the phylogenetic analysis based on core genes, we next computed ANIs of our cryptic
248 C-II isolates versus all other genomes (Fig 1B and C). This analysis shows that our isolates have ANIs
249 of > 98.7% against all other C-II isolates, but that ANI values to other clades fall well below the 96%
250 that has been named as a cut-off for the same species(34).

251 These data demonstrate that *C. difficile* ribotype 151 is polyphyletic with isolates from the
252 same ribotype falling into two different cryptic clades as well as a 'regular' clade, which has not been
253 reported before. Though there is good concordance between PCR ribotype and MLST-based
254 approaches(13, 35, 36), these findings underscore that PCR ribotyping is insufficiently discriminatory
255 for phylogenetic placement.

256

257 **A pan-genome analysis suggests that accessory genes are conserved within ST205**

258 It is conceivable that despite divergent core genomes, RT151 share a common accessory genome.
259 Therefore, we next investigated the accessory genes. We identified a total of 14,297 accessory genes
260 and these were clustered using the complete linkage method with Euclidian distances in the
261 ComplexHeatmap package(27) to identify potential similarities in the accessory genome of RT151 or

262 ST205 (Figure 2). This analysis largely confirms the core genome phylogeny (Figure 1), as a clear
263 clustering based on accessory genes is seen within RT151 isolates belonging to ST205, while the
264 cryptic clade isolates from RT151 have divergent accessory genomes. In addition, isolates from the
265 same clade also generally cluster together, with the exception of clade 2 isolates falling within the
266 larger clade 1 cluster.

267

268 **Cryptic clade C-II RT151 isolates can have a mono-toxin pathogenicity locus at a novel insertion**
269 **site**

270 On the basis of our initial Illumina-sequenced genomes we were unable to recover PaLoc sequences
271 on a single contig for all isolates and, importantly, genomic neighborhoods around the PaLoc could
272 not be fully resolved. In particular, for the *tcdB*-positive C-II isolate from the Netherlands
273 (LUMCMM21 0003), *tcdB* was contained on a 18-kb contig lacking flanking chromosomal regions.
274 Additionally, we noted limited homology to the pHSJD-312 plasmid (Supplemental Figure S2A)(8),
275 which carries a similar PaLoc. However, in contrast to the PaLoc on this plasmid, the LUMCMM21
276 0003 PaLoc lacked the binary toxin genes, *cdtA* and *cdtB*.

277 To identify the PaLoc insertion site and to establish whether it is carried on plasmid, we
278 generated complete genomes for the three cryptic clade RT151 isolate using PacBio circular
279 consensus sequencing (Table 1). The non-toxigenic C-I isolate (LUMCMM16 0013) was found to
280 contain extrachromosomal elements (ECEs) of 132 kb and 60 kb, that are classified in a PHASTER
281 prediction as possible and putative phage, respectively. The *tcdB* positive C-II isolate (LUMCMM21
282 0003) showed only a single contig, demonstrating that the toxin gene is not carried on a plasmid.
283 The *tcdA*-positive C-II isolate (LUMCMM19 2333) harbors a 13-kb ECE plasmid but this plasmid does
284 not contain toxin genes. Thus, in contrast to previous reports from C-I strains(4, 8), the cryptic RT151
285 isolates appear to carry their toxins on the chromosome. To further classify our toxin genes, we used

286 DiffBase(37) and this revealed LUMCMM21 0003 to have tcdB subtype B12, while LUMCMM19 2333
287 had toxin subtype A7.

288 PacBio sequencing also allows us to identify motifs associated with m4C and m6A
289 methylation events (Supplemental Table S2). Methylation has been reported to affect virulence
290 characteristics of *C. difficile*(38). Notably, all strains contain the CamA-dependent modification of
291 CAAAA**A** sequences (modified residue in bold)(38, 39), indicating that the action of this
292 methyltransferase is conserved in the RT151 C-I and C-II strains. We also observed m6A methylation
293 on a subset of CAAA**A**AASNV motifs in the C-II strains, but not the C-I strain, suggesting it may be
294 specific for clade C-II. Finally, we observed m4C modification on a subset of **C**TATTATCW motifs in
295 the C-II strains, that may be conserved in C-I strains (CWATTATCW). This motif has not previously
296 been identified by others(38).

297 Next, we investigated the genomic location of the PaLoc's of the RT151 isolates by aligning
298 the observed insertion sites to those known ones from literature (Figure 1A)(4, 6), using the genome
299 sequence of the RT012 reference strain 630 and the sequence of plasmid pHSJD-312as a reference
300 (Figure 3)(4, 33). This analysis reveals that the PaLoc in the Dutch C-II isolate is inserted at a location
301 that has not previously been described, between genes homologous to *cd0628* and *cd0629* of strain
302 630 (Figure 3C).

303 The PaLoc of LUMCMM21 0003, as well as the *cdu1-cdd1* insertion site that normally
304 harbors the PaLoc, is reminiscent of what has been described for a toxin A-negative toxin B-positive
305 toxinotype XXXII strain, 173070(5). Similar to our isolate, the *cdu1-cdd1* locus contains a multi-gene
306 insertion interrupted by an IS256 insertion element and the PaLoc has a *tcdR-tcdB-tcdE* architecture,
307 lacks the binary toxin genes but has a divergently transcribed transcriptional regulator gene
308 upstream from *tcdR* (presumably this is *cdtR*). By downloading raw reads (SRR1514909) and
309 performing ANI analyses on the assembled genome, we confirmed that 173070 was also a cryptic C-
310 II isolate (data not shown). This is in line with the SLO148 ribotype reported in the manuscript(5) to

311 be highly similar to that of RT151 (data not shown) and the reported ST (ST200) is classified as C-II in
312 other studies(2, 40).

313 The PaLoc in the Maltese C-II isolate (LUMCMM19 2333) is inserted between *spoVAE* and
314 *cd0776*, as previously described for isolate RA09-70(6) (Figure 1A, Supplemental Figure S2). RA09-70
315 was classified as Clade 5 but the phylogenetic analysis by the authors of the manuscript included
316 only strains from clade 1-5 and C-I. We therefore downloaded the respective genome assembly
317 (GCF_001299495.1) and confirmed through ANI analyses that RA09-70 is, in fact, a clade C-II isolate
318 that was misclassified as clade 5 for lack of reference (data not shown).

319

320 **The role of C-II isolates in CDI**

321 To the best of our knowledge, the present study is the first to conclusively link a clade C-II isolate to
322 human disease, although the above sentences indicate that clade C-II isolates could be more
323 common than hitherto assumed. Nevertheless, diagnosis of CDI due to LUMCMM21 0001/0003 was
324 hampered by the fact that fecal samples, as well as *C. difficile* isolates, tested negative in the Xpert *C.*
325 *difficile* BT assay, used by many laboratories worldwide. Toxin gene sequences of cryptic clade
326 isolates can differ significantly from those of regular clade 1-5 isolates which can lead to false
327 negative results(4, 5). Due to the proprietary nature of the Xpert *C. difficile* BT assay, we could not
328 confirm that this failure is due to mismatches between the primers used in the assay and the C-II
329 *tcdB* sequence but we have informed the company and provided the WGS data to investigate this
330 further. We do note that positive identifications were made using the ECDC-endorsed multiplex PCR
331 (relating to little or no mismatches in the primer sequence compared to the genome). A recent
332 report investigating *C. difficile* isolates from environmental locations showed several atypical PaLoc
333 locations in cryptic clade isolates and *in silico* PCR further suggested that routine diagnostic assays
334 would return highly variable results if such isolates would be encountered in the clinic(3), which is in

335 line with what we found in the current study. This underscores the potential of these isolates to be
336 OneHealth pathogens. Toxin-based assays (CCNA and VIDAS®) demonstrated toxin production for
337 the mono-toxin B isolate (LUMCMM21 0001/0003), but not the mono-toxin A isolate (LUMCMM19
338 2333). Though the latter was isolated from a hospitalized patient, lack of metadata did not allow us
339 to establish whether the patient had symptoms consistent with CDI and it remains to be established
340 if this isolate is pathogenic. Recently, it has been reported that toxin A-positive toxin B-negative C-III
341 strains cannot be identified on ChromID agar due to their esculin-negative nature(41). We found,
342 however, that all RT151 strains – including those belonging to clades C-I and C-II, were esculin-
343 positive (Supplemental Table S1).

344 Altogether, reports of cryptic clade isolates that cause human CDI are still very rare. With
345 the increasing rate of community-onset CDI(42) and the discovery that cryptic clade isolates with
346 atypical PaLoc configurations are widespread in the environment (3), it is very likely that other CDI
347 cases caused by isolates from this cryptic clade occur but are missed by standard diagnostics if only
348 PCR is used.

349

Acknowledgements

The authors thank K. Galea (L-Università ta' Malta); J. Freeman and K. Davies (University of Leeds) for sharing unpublished data; Michael Perry and Trefor Morris (Anaerobic Reference Unit, Cardiff) for providing RT151 whole genome sequence information; I. Sidorov for implementation of clinker; and members of the Experimental Bacteriology Group at the LUMC for helpful discussions.

Funding

No external funding was received for this study.

Conflict of interest

EJK holds an unrestricted research grant from Vedanta Biosciences. All other authors: none to declare.

Author contributions

QRD: software, validation, formal analysis, data curation, writing - original draft, visualisation; JTvdB: writing – original draft; CH: investigation ; IMJGS: investigation; LD: resources; ACF: investigation; RHAMV: investigation, formal analysis; SLK: resources, supervision; EJK: conceptualization, writing - original draft; WKS: conceptualization, formal analysis, writing - original draft, visualization, supervision. All authors reviewed and edited the manuscript, and approved the final version.

369 **LEGENDS**

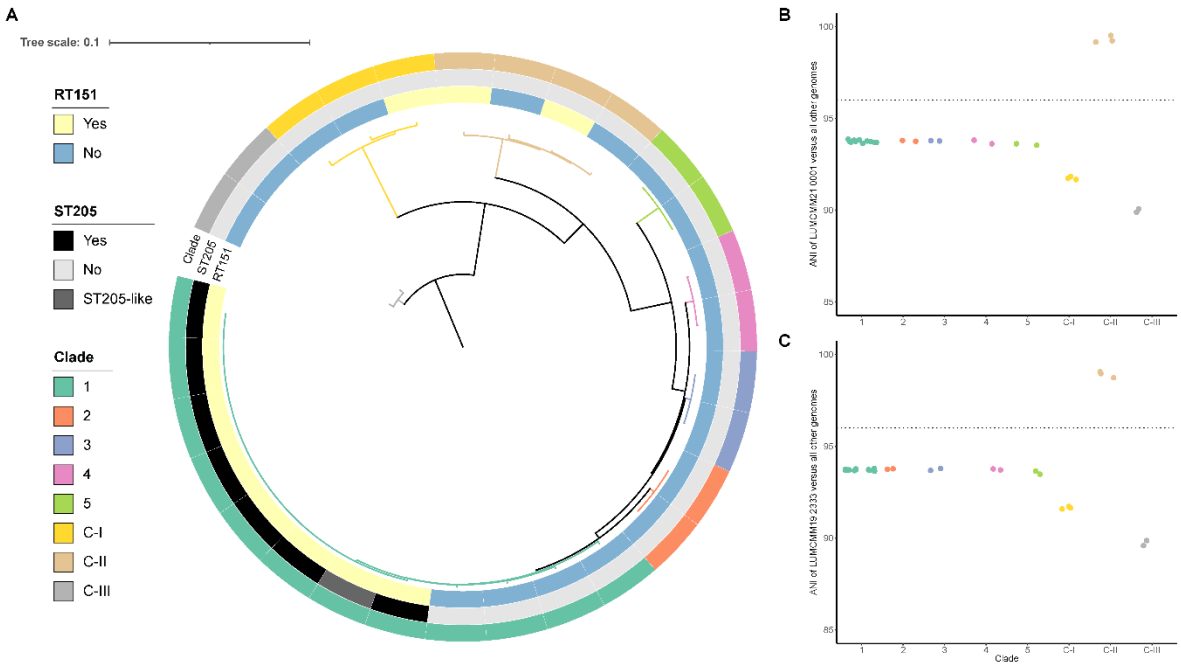
370 **Figure 1. RT151 isolates are part of clade 1, cryptic clade C-I and C-II and belong to various**

371 **sequence types**. **A.** The tree was rooted at midpoint in iTOL version 6. Detailed information,

372 including the depicted classifications, are available in Supplemental Table S1. **B.** ANI calculations of

373 LUMCMM21 0001/0003 versus all other included isolate genomes. **C.** ANI calculations of

374 LUMCMM19 2333 versus all other included isolate genomes.



375

376

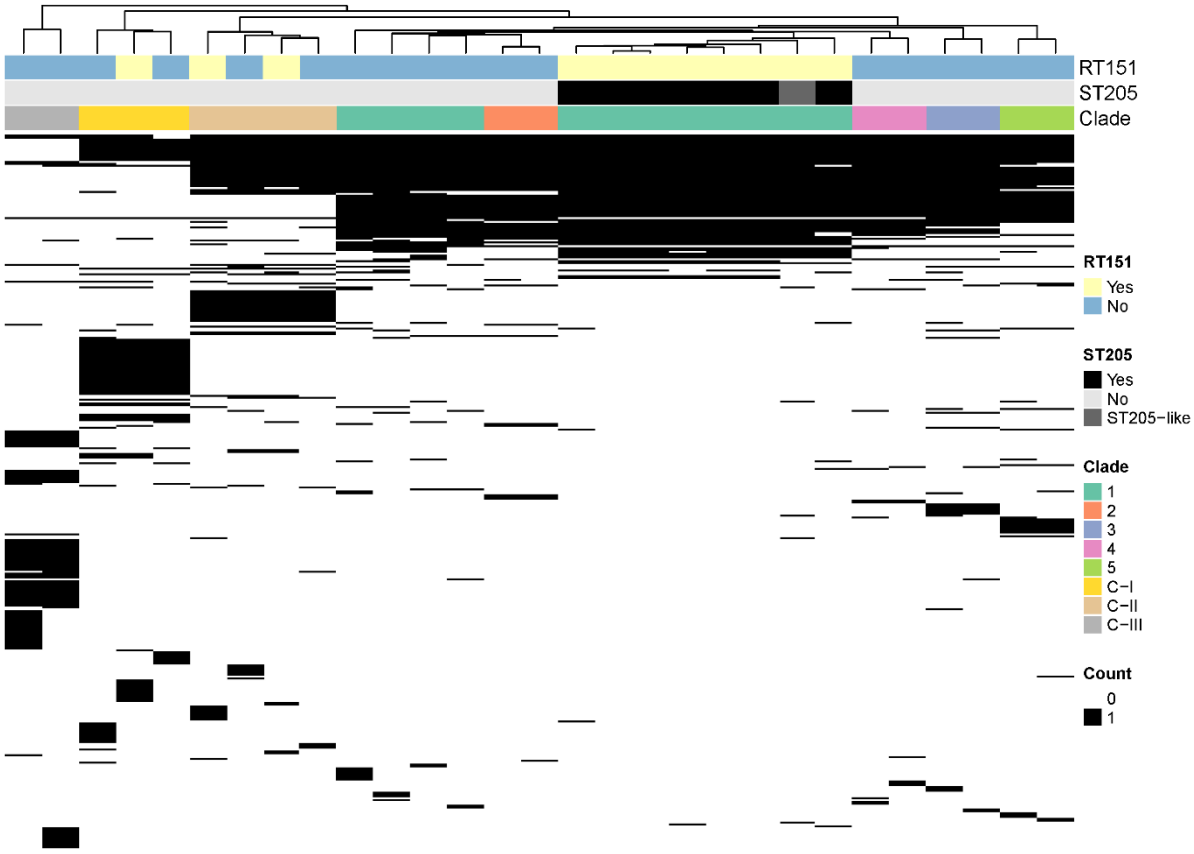
377 **Figure 2. No clustering of RT151 isolates based on the accessory genome (n = 14,297 genes).** Black

378 denotes the presence of a gene, white denotes the absence of a gene. Clustering of accessory genes

379 was performed using the complete linkage method with Euclidian distances in the ComplexHeatmap

380 package(27). Detailed information, including the depicted classifications, are available in

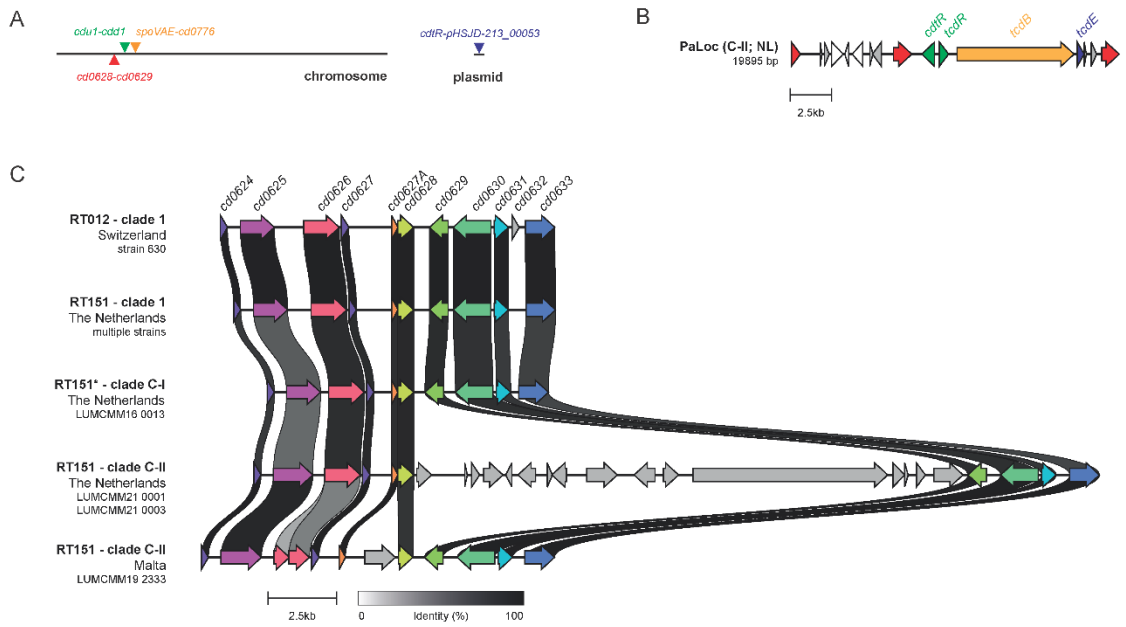
381 Supplemental Table S1. Coloring is the same as in Figure 1.



382

383

384 **Figure 3. Structure of the mono-toxin pathogenicity locus and PaLoc insertion sites of two cryptic**
385 **clade-II RT151 strains. A.** Overview of known PaLoc insertion sites (above the lines)(4, 6) and the
386 novel PaLoc insertion site (below the line) identified in the present study. Three sites are located on
387 the chromosome (nomenclature from RT012 reference strain 630)(33), and one is located on a
388 plasmid (nomenclature from clade C-I strain strain HSJD-312)(4, 8). **B**. Structure of the PaLoc in the
389 A-B+ RT151 clade C-II strain (LUMCMM21 0001/0003) showing all open reading frames predicted by
390 Prokka(24). Genes involved in transposition or recombination are indicated in red, hypothetical
391 genes with homology to the PaLoc on pHSJD-312 are indicated in gray, genes encoding toxin
392 regulators are indicated in green, the toxin gene is in orange and the holin gene in purple. **C**.
393 Alignment of the chromosomal region incorporating the *cd0628-cd0629* insertion site for RT151
394 isolates in comparison with the RT012 reference strain 630(33). RT151 clade 1 is LUMCMM11 0004,
395 RT151 clade C-I is LUMCMM16 0013, and RT151 clade C-II are LUMCMM21 0001/0003 (The
396 Netherlands) and LUMCMM19 2333 (Malta). Figure generated by clinker(43)and centered on
397 *cd0628*. Colored arrows indicate similar genes; links are drawn between similar genes on
398 neighboring clusters and are shaded based on sequence identity (0% white, 100% black, identity
399 threshold for visualization 0.35).



400

401 **Table 1. Genome characteristics of the cryptic clade RT151 isolates**

Isolate	Clade	Size (bp)	%GC	#ORFs	GenBank
LUMCMM16 0013	C-I	4169505 (chromosome) 132414 (possible phage) 59572 (putative phage)	28.8% 26.3% 26.6%	3747 174 79	GCA_945909635
LUMCMM21 0003	C-II	4108918 (chromosome)	28.7%	3688	GCA_945861085
LUMCMM19 2333	C-II	4424808 (chromosome) 12730 (putative plasmid)	28.7% 25.9%	4020 17	GCA_945909465

402

403 **REFERENCES**

404 1. Smits WK, Lyras D, Lacy DB, Wilcox MH, Kuijper EJ. Clostridium difficile infection. Nat Rev Dis
405 Primers. 2016;2:16020.
406 2. Knight DR, Imwattana K, Kullin B, Guerrero-Araya E, Paredes-Sabja D, Didelot X, et al. Major
407 genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. eLife. 2021;10.
408 3. Williamson CHD, Stone NE, Nunnally AE, Roe CC, Vazquez AJ, Lucero SA, et al. Identification
409 of novel, cryptic Clostridioides species isolates from environmental samples collected from diverse
410 geographical locations. Microbial genomics. 2022;8(2).
411 4. Ramírez-Vargas G, López-Ureña D, Badilla A, Orozco-Aguilar J, Murillo T, Rojas P, et al. Novel
412 Clade C-I Clostridium difficile strains escape diagnostic tests, differ in pathogenicity potential and
413 carry toxins on extrachromosomal elements. Scientific reports. 2018;8(1):13951.
414 5. Janezic S, Marín M, Martín A, Rupnik M. A new type of toxin A-negative, toxin B-positive
415 Clostridium difficile strain lacking a complete tcdA gene. Journal of clinical microbiology.
416 2015;53(2):692-5.
417 6. Monot M, Eckert C, Lemire A, Hamiot A, Dubois T, Tessier C, et al. Clostridium difficile: New
418 Insights into the Evolution of the Pathogenicity Locus. Scientific reports. 2015;5:15023.
419 7. Braun V, Hundsberger T, Leukel P, Sauerborn M, von Eichel-Streiber C. Definition of the
420 single integration site of the pathogenicity locus in Clostridium difficile. Gene. 1996;181(1-2):29-38.
421 8. Ramírez-Vargas G, Rodríguez C. Putative Conjugative Plasmids with tcdB and cdtAB Genes in
422 Clostridioides difficile. Emerging infectious diseases. 2020;26(9):2287-90.
423 9. Fawley WN, Knetsch CW, MacCannell DR, Harmanus C, Du T, Mulvey MR, et al. Development
424 and validation of an internationally-standardized, high-resolution capillary gel-based electrophoresis
425 PCR-ribotyping protocol for Clostridium difficile. PloS one. 2015;10(2):e0118150.
426 10. Control ECfDPa. Laboratory procedures for diagnosis and typing of human *Clostridium*
427 *difficile* infection. Stockholm: ECDC; 2018.
428 11. Wolfhagen MJ, Fluit AC, Torensma R, Poppelier MJ, Verhoef J. Rapid detection of toxigenic
429 Clostridium difficile in fecal samples by magnetic immuno PCR assay. Journal of clinical microbiology.
430 1994;32(7):1629-33.
431 12. van den Berg RJ, Bruijnesteijn van Coppenraet LS, Gerritsen HJ, Endtz HP, van der Vorm ER,
432 Kuijper EJ. Prospective multicenter evaluation of a new immunoassay and real-time PCR for rapid
433 diagnosis of Clostridium difficile-associated diarrhea in hospitalized patients. Journal of clinical
434 microbiology. 2005;43(10):5338-40.
435 13. Baktash A, Corver J, Harmanus C, Smits WK, Fawley W, Wilcox MH, et al. Comparison of
436 Whole-Genome Sequence-Based Methods and PCR Ribotyping for Subtyping of Clostridioides
437 difficile. Journal of clinical microbiology. 2022;60(2):e0173721.
438 14. Knetsch CW, Terveer EM, Lauber C, Gorbalenya AE, Harmanus C, Kuijper EJ, et al.
439 Comparative analysis of an expanded Clostridium difficile reference strain collection reveals genetic
440 diversity and evolution through six lineages. Infect Genet Evol. 2012;12(7):1577-85.
441 15. Chen S, Zhou Y, Chen Y, Gu J. fastp: an ultra-fast all-in-one FASTQ preprocessor.
442 Bioinformatics (Oxford, England). 2018;34(17):i884-i90.
443 16. Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham
444 Bioinformatics, Babraham Institute, Cambridge, United Kingdom; 2010.
445 17. Ewels P, Magnusson M, Lundin S, Käller M. MultiQC: summarize analysis results for multiple
446 tools and samples in a single report. Bioinformatics (Oxford, England). 2016;32(19):3047-8.
447 18. Souvorov A, Agarwala R, Lipman DJ. SKESA: strategic k-mer extension for scrupulous
448 assemblies. Genome biology. 2018;19(1):153.
449 19. Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. Using SPAdes De Novo
450 Assembler. Current protocols in bioinformatics. 2020;70(1):e102.
451 20. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome
452 assemblies. Bioinformatics (Oxford, England). 2013;29(8):1072-5.

453 21. Kolmogorov M, Yuan J, Lin Y, Pevzner PA. Assembly of long, error-prone reads using repeat
454 graphs. Nat Biotechnol. 2019;37(5):540-6.
455 22. Hunt M, Silva ND, Otto TD, Parkhill J, Keane JA, Harris SR. Circlator: automated circularization
456 of genome assemblies using long sequencing reads. Genome biology. 2015;16:294.
457 23. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and
458 Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of
459 Eukaryotic, Prokaryotic, and Viral Genomes. Molecular biology and evolution. 2021;38(10):4647-54.
460 24. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics (Oxford, England).
461 2014;30(14):2068-9.
462 25. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MT, et al. Roary: rapid large-scale
463 prokaryote pan genome analysis. Bioinformatics. 2015;31(22):3691-3.
464 26. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence
465 alignment based on fast Fourier transform. Nucleic acids research. 2002;30(14):3059-66.
466 27. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in
467 multidimensional genomic data. Bioinformatics (Oxford, England). 2016;32(18):2847-9.
468 28. Hoang DT, Chernomor O, von Haeseler A, Minh BQ, Vinh LS. UFBoot2: Improving the
469 Ultrafast Bootstrap Approximation. Molecular biology and evolution. 2018;35(2):518-22.
470 29. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-
471 TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era.
472 Molecular biology and evolution. 2020;37(5):1530-4.
473 30. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree
474 display and annotation. Nucleic acids research. 2021;49(W1):W293-w6.
475 31. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb
476 software, the PubMLST.org website and their applications. Wellcome Open Res. 2018;3:124.
477 32. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI
478 analysis of 90K prokaryotic genomes reveals clear species boundaries. Nature Communications.
479 2018;9(1):5114.
480 33. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, Stabler R, et al. The multidrug-
481 resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nature genetics.
482 2006;38(7):779-86.
483 34. Richter M, Rosselló-Móra R. Shifting the genomic gold standard for the prokaryotic species
484 definition. Proc Natl Acad Sci U S A. 2009;106(45):19126-31.
485 35. Zhao H, Nickle DC, Zeng Z, Law PYT, Wilcox MH, Chen L, et al. Global Landscape of
486 Clostridioides Difficile Phylogeography, Antibiotic Susceptibility, and Toxin Polymorphisms by Post-
487 Hoc Whole-Genome Sequencing from the MODIFY I/II Studies. Infect Dis Ther. 2021;10(2):853-70.
488 36. Seth-Smith HMB, Biggel M, Roloff T, Hinic V, Bodmer T, Risch M, et al. Transition From PCR-
489 Ribotyping to Whole Genome Sequencing Based Typing of Clostridioides difficile. Front Cell Infect
490 Microbiol. 2021;11:681518.
491 37. Mansfield MJ, Tremblay BJ, Zeng J, Wei X, Hodgins H, Worley J, et al. Phylogenomics of 8,839
492 Clostridioides difficile genomes reveals recombination-driven evolution and diversification of toxin A
493 and B. PLoS Pathog. 2020;16(12):e1009181.
494 38. Oliveira PH, Ribis JW, Garrett EM, Trzilova D, Kim A, Sekulovic O, et al. Epigenomic
495 characterization of Clostridioides difficile finds a conserved DNA methyltransferase that mediates
496 sporulation and pathogenesis. Nat Microbiol. 2020;5(1):166-80.
497 39. van Eijk E, Anvar SY, Browne HP, Leung WY, Frank J, Schmitz AM, et al. Complete genome
498 sequence of the Clostridium difficile laboratory strain 630Δerm reveals differences from strain 630,
499 including translocation of the mobile element CTn5. BMC Genomics. 2015;16(1):31.
500 40. Imwattana K, Knight DR, Kullin B, Collins DA, Putsathit P, Kiratisin P, et al. Clostridium difficile
501 ribotype 017 - characterization, evolution and epidemiology of the dominant strain in Asia. Emerg
502 Microbes Infect. 2019;8(1):796-807.

503 41. Shivaperumal N, Knight DR, Imwattana K, Androga GO, Chang BJ, Riley TV. Esculin hydrolysis
504 negative and TcdA-only producing strains of clostridium (Clostridiodes) difficile from the
505 environment in Western Australia. Journal of applied microbiology. 2022.
506 42. Turner NA, Grambow SC, Woods CW, Fowler VG, Jr., Moehring RW, Anderson DJ, et al.
507 Epidemiologic Trends in Clostridioides difficile Infections in a Regional Community Hospital Network.
508 JAMA Netw Open. 2019;2(10):e1914149.
509 43. Gilchrist CLM, Chooi Y-H. clinker & clustermap.js: automatic generation of gene cluster
510 comparison figures. Bioinformatics (Oxford, England). 2021;37(16):2473-5.

511

512

SUPPLEMENTAL INFORMATION

Supplemental Table S1. Characteristics of RT151 isolates and reference sequences analyzed in this study.

PCR RT151 isolates

Name	Clade	RT	RT151	ST	ST205[1]	Accession_Illumina_reads	Accession_Pacbio	Accession_Assembly	Sample_alias	Identifier	Reference	Country_Isolation	State_city	Source	Year_Isolation	Collection	Xpert_Cdifficile_BT	Toxin_neutralization_assay	VIDAS	LUMC_multiplex[3]	UMCU_tcdB_RTPCR[4]	ChromID
R18152B	1	151	Yes	948	ST205-like	ERR9902328			Cdiff_RT151_1	R18152B	This_study	Germany	Freiburg		2003	UK Anaerobic Reference Unit						
R18284B	1	151	Yes	205	Yes	ERR9902329			Cdiff_RT151_2	R18284B	This_study	USA	Arizona		2003	UK Anaerobic Reference Unit						
R20189B	1	151	Yes	205	Yes	ERR9902330			Cdiff_RT151_3	R20189B	This_study	USA	Arizona		2004	UK Anaerobic Reference Unit						
R54230B	1	151	Yes	205	Yes	ERR9902331			Cdiff_RT151_4	R54230B	This_study	Wales	Newport		2010	UK Anaerobic Reference Unit						
LUMCMM11 0004	1	151	Yes	205	Yes	ERR9902332			Cdiff_RT151_5	104875-001-011	This_study	Netherlands	Nijmegen		2011	LUMC	negative	negative	negative	11000	ND	positive
LUMCMM16 0013	C-I	151	Yes	759	No	ERR9902333	ERR9902350	GCA_945909635	Cdiff_RT151_6	104875-001-012	This_study	Netherlands	Nijmegen		2016	LUMC	negative	negative	equivalent	11000	ND	positive
LUMCMM21 0001 / LUMCMM21 0003	C-II	151	Yes	946	No	ERR9902334	ERR9902351	GCA_945863085	Cdiff_RT151_7	104875-001-013	This_study	Netherlands	Utrecht	CDI_patient	2021	LUMC	negative	weak[2]	positive[2]	11000	positive	positive
LUMCMM19 2333	C-II	151	Yes	947	No	ERR9902335	ERR9902352	GCA_945909465	Cdiff_RT151_8	104875-001-016	This_study	Malta			2019	LUMC	negative	negative	equivalent	11100	ND	positive
LUMCMM19 2334	1	151	Yes	205	Yes	ERR9902336			Cdiff_RT151_9	104875-001-017	This_study	Greece			2019	LUMC	negative	negative	equivalent	11000	ND	positive
LUMCMM16 0014	1	151	Yes	205	Yes	ERR9902337			Cdiff_RT151_10	104875-001-018	This_study	Spain		Animal	2016	LUMC	negative	negative	negative	11000	ND	positive
LUMCMM16 0015	1	151	Yes	205	Yes	ERR9902338			Cdiff_RT151_11	104875-001-019	This_study	Spain		Animal	2016	LUMC	negative	negative	equivalent	11000	ND	positive
Reference sequences																						
	C-III	Unknown	No	369	No	ERR2215981				ERR2215981	10.1128/ICM.01987-17											
	C-III	Unknown	No	343	No	ERR2216003				ERR2216003	10.1128/ICM.01987-17											
	C-II	Unknown	No	200	No	SRR3654506				SRR3654506	10.1371/journal.pone.0182307											
	C-II	Unknown	No	637	No	ERR3296451				ERR3296451	10.1038/s41598-019-43464-4											
	C-I	Unknown	No	178	No	ERR232399				ERR232399	10.1093/gbe/evz204											
	C-I	Unknown	No	314	No	SRR3629659				SRR3629659	10.1371/journal.pone.0182307											
	1	1	No	3	No	SRR1519369				001-LL-SRR1519369	10.1128/ICM.01737-21											
	1	12	No	54	No	SRR1519379				012-LL-SRR1519379	10.1128/ICM.01737-21											
	2	16	No	1	No	SRR1519382				016-LL-SRR1519382	10.1128/ICM.01737-21											
	4	17	No	37	No	SRR1519383				017-LL-SRR1519383	10.1128/ICM.01737-21											
	3	23	No	5	No	SRR1519388				023-LL-SRR1519388	10.1128/ICM.01737-21											
	1	26	No	7	No	SRR1519390				026-LL-SRR1519390	10.1128/ICM.01737-21											
	2	27	No	1	No	SRR1519391				027-LL-SRR1519391	10.1128/ICM.01737-21											
	4	40	No	121	No	SRR1519398				040-LL-SRR1519398	10.1128/ICM.01737-21											
	1	57	No	55	No	SRR1519410				057-LL-SRR1519410	10.1128/ICM.01737-21											
	3	58	No	96	No	SRR1519411				058-LL-SRR1519411	10.1128/ICM.01737-21											
	5	66	No	11	No	SRR1519416				066-LL-SRR1519416	10.1128/ICM.01737-21											
	5	78	No	11	No	SRR1519423				078-LL-SRR1519423	10.1128/ICM.01737-21											

Abbreviations: RT = PCR ribotype, ST = (7-gene multilocus) sequence type

[1] classifier used in Figures 1 and 2

[2] "weak" indicates a partial cytopathic effect, only observed with undiluted culture supernatant, on VeroC6 cells after 48h

[3] "positive" corresponds to a value of 0.99

[4] Numbers indicate positive (1) or negative (0) for 16S, gluD, tcdA, tcdB and cdt (in that order). For reference see main manuscript.

Supplemental Table S2. Overview of modified base motifs of the cryptic clade RT151 isolates.

LUMCMM16 0013 modified base motifs

Motif	Modified Position	Modification Type	% of Motifs Detected	# of Motifs Detected	# of Motifs in Genome	Mean QV	Mean Coverage	Partner Motif	Mean IPD ratio	Group Tag	Objective Score
CAAAAA	6	m6A	99.7%	7,75	7,767	323.6	107.5		4.2	CAAAAA	2,502,921.2
CTKMAG	5	m6A	70.1%	3,785	5,398	359.9	109.2	CTKMAG	5.1	CTKMAG	989,662.4
VNCYSVNYAD	3	m4C	22.3%	1,897	8,475	61.9	111.3		1.7	VNCYSVNYAD	30,515.8
GNGGYVSCRN$_n$D	3	unknown	51.7%	118	228	60.6	115.8		1.8	GNGGYVSCRNNNNND	3,950.7
CGN$_2$TNGNGACNNANC	11	unknown	96.0%	24	25	68.6	126.3		1.8	CGNNTNGNGACNNANC	1,586.6
DN$_2$ACTGAAGR	9	m6A	24.5%	52	212	94.2	109.8		2.2	DNNNNNNNACTGAAGR	1,227.8
GGNCGTN$_2$CNCTN$_4$T	4	unknown	100.0%	14	14	82.5	125.3		1.9	GGNCGTNNCNCTNNNNT	1,155.0
HTTATACNWN$_4$C	3	unknown	15.2%	117	769	51.9	105.7		1.6	HTTATACNWNNNNC	1,115.4
TN$_2$CNGANCNCNCN$_2$TNA	4	unknown	95.8%	23	24	49.9	126.7		1.6	TNNCNGANCNCNCNNTNA	1,103.9
CWATTATCW	1	m4C	17.2%	93	538	50.8	109.7		1.6	CWATTATCW	972.7
GGN$_2$CGAMATT	2	unknown	78.5%	22	28	45.8	126.6		1.6	GGNNNCGAMATT	810.5
AN$_2$TCGNCTNANCNC	8	unknown	100.0%	13	13	59.0	125.0		1.6	ANNTCGNCTNANCNC	767.0
TN$_2$CGGNGGN$_2$GTA	9	unknown	86.6%	13	15	59.6	124.8		1.7	TNNNCGGNGGNNGTA	681.3
CGNCCTN$_2$CAN$_2$GC	5	m4C	100.0%	12	12	53.5	126.8		1.6	CGNCCTNNNCANNNGC	642.0

LUMCMM21 0003 modified base motifs

Modified Base Motifs

Motif	Modified Position	Modification Type	% of Motifs Detected	# of Motifs Detected	# of Motifs in Genome	Mean QV	Mean Coverage	Partner Motif	Mean IPD ratio	Group Tag	Objective Score
CAAAAA	6	m6A	98.2%	7,668	7,807	478.3	185.5		4.1	CAAAAA	3,608,910.0
VNCYGRNCRC	3	unknown	55.1%	70	127	74.3	251.6		1.6	VNCYGRNCRC	3,042.6
CAAAAAASNV	5	m6A	16.5%	102	617	148.9	185.8		2.1	CAAAAAASNV	3,006.4
HN$_2$GCVDGCNHC	5	m4C	40.5%	73	180	62.7	226.2		1.6	HNNGCVDGCNHC	2,032.4
AN$_2$CGNCTNMNCAC	8	unknown	100.0%	24	24	68.3	314.3		1.6	ANNNCGNCTNMNCAC	1,639.0
CTATTATCW	1	m4C	28.4%	84	295	58.8	179.5		1.6	CTATTATCW	1,593.6
GTCN$_2$ACANCN$_2$CG	1	unknown	100.0%	12	12	53.6	314.0		1.5	GTCNNNACANCNNNNNCG	643.0

LUMCMM19 2333 modified base motifs

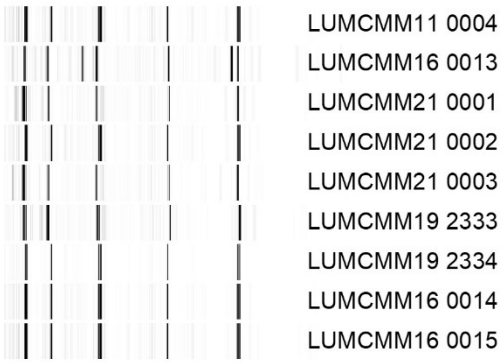
Modified Base Motifs

Motif	Modified Position	Modification Type	% of Motifs Detected	# of Motifs Detected	# of Motifs in Genome	Mean QV	Mean Coverage	Partner Motif	Mean IPD ratio	Group Tag	Objective Score
CAAAAA	6	m6A	99.5%	8,244	8,285	415.2	165.7		4.0	CAAAAA	3,407,425.2
CCAAAT	5	m6A	100.0%	3,912	3,912	489.2	166.0		4.3	CCAAAT	1,913,557.0
TAGN$_2$CTG	2	m6A	100.0%	1,175	1,175	597.9	172.2	CAGNNNNNCTA	5.7	CAGNNNNNCTA/TAGNNNNNCTG	702,564.0
CAGN$_2$CTA	2	m6A	100.0%	1,175	1,175	588.2	172.2	TAGNNNNNCTG	5.6	CAGNNNNNCTA/TAGNNNNNCTG	691,190.0
RGAKCT	3	m6A	12.7%	554	4,36	75.7	167.9		1.9	RGAKCT	6,547.8
CAAAAAASNV	5	m6A	15.0%	98	652	142.6	167.4		2.1	CAAAAAASNV	2,537.9
CTATTATCW	1	m4C	23.4%	77	329	53.1	164.9		1.6	CTATTATCW	1,107.4
CN$_2$TNAAN$_2$GCGGC	13	unknown	100.0%	13	13	83.8	243.3		1.7	CNNNTNAANNGCGGC	1,089.0
TN$_2$TN$_2$GGCNCANCA	11	unknown	86.3%	19	22	68.2	234.8		1.6	TNNNTNNNGGCNCANCA	1,013.4

Supplemental Table S3. Overview of MLST profiles of all RT151 isolates

Name	Clade	RT	ST	adk	atpA	dxr	glyA	recA	sodA	tpi	Closest_ST	Number_Identical_Alleles_MLST_to_closest_ST
R18152B	1	151	948	1	3	13	1	5	116	1	205	6
R18284B	1	151	205	1	3	13	1	5	23	1		
R20189B	1	151	205	1	3	13	1	5	23	1		
R54230B	1	151	205	1	3	13	1	5	23	1		
LUMCMM11 0004	1	151	205	1	3	13	1	5	23	1		
LUMCMM16 0013	C-I	151	759	13	18	22	105	63	32	28		
LUMCMM21 0001 / LUMCMM21 0003	C-II	151	946	15	21	28	38	82	106	30	200	5
LUMCMM19 2333	C-II	151	947	53	21	58	140	48	36	30	600	1
LUMCMM19 2334	1	151	205	1	3	13	1	5	23	1		
LUMCMM16 0014	1	151	205	1	3	13	1	5	23	1		
LUMCMM16 0015	1	151	205	1	3	13	1	5	23	1		

521 **Supplemental Figure S1. PCR RT banding profiles of all RT151 isolates.**



522

523 **Supplemental Figure S2.** Alignment of previously characterized PaLoc insertion sites for RT151

524 isolates in comparison with the RT012 reference strain 630(33). **A**. Extrachromosomal element

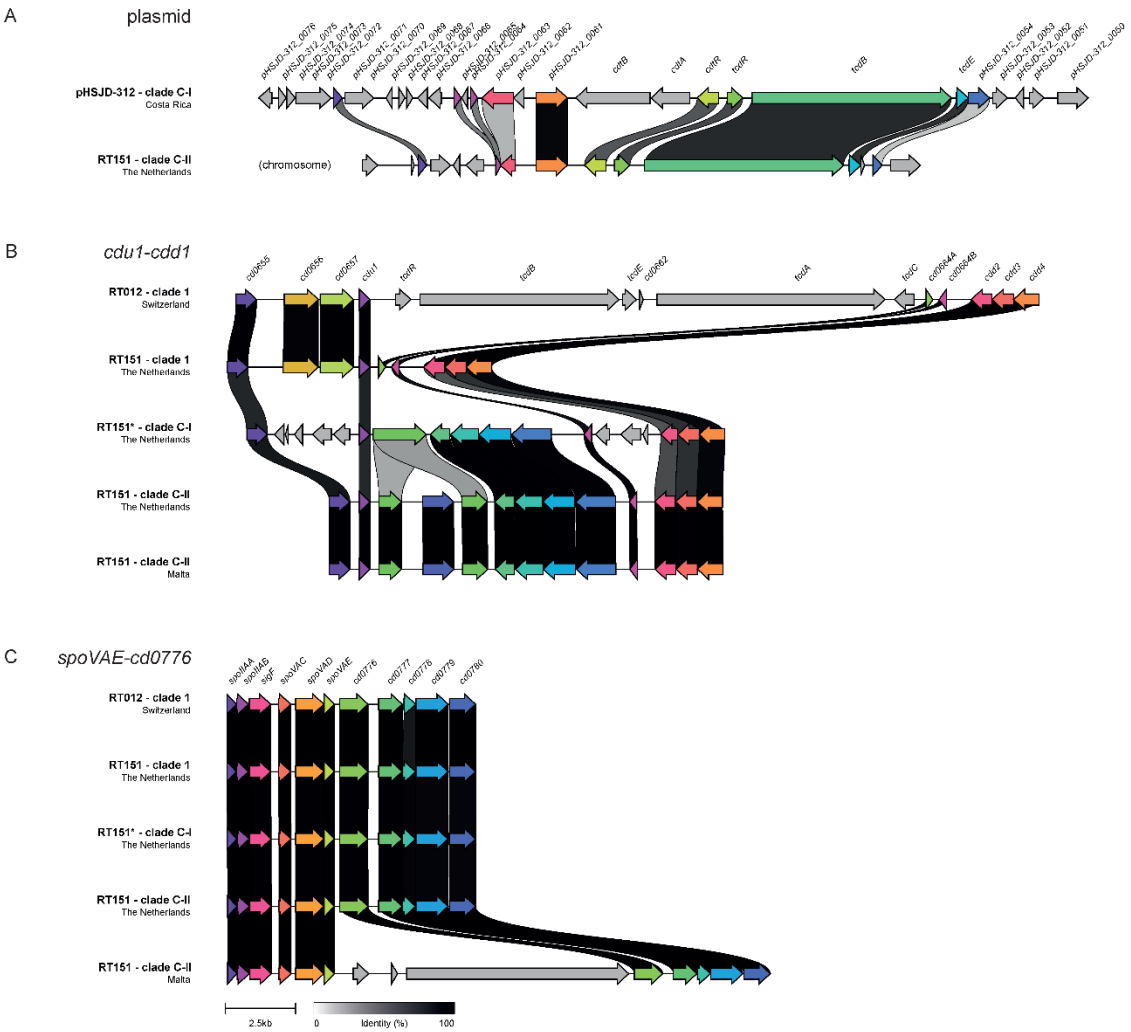
525 pHSJD-312. **B**. *cdu1-cdd1*. **C**. *spoVAE-cd0776*)(6). RT151 clade 1 is LUMCMM11 0004, RT151 clade C-I

526 is LUMCMM16 0013, and RT151 clade C-II are LUMCMM21 0001/0003 (The Netherlands) and

527 LUMCMM19 2333 (Malta). Figure generated by clinker(43). Colored arrows indicate similar genes;

528 links are drawn between similar genes on neighboring clusters and are shaded based on sequence

529 identity (0% white, 100% black, identity threshold for visualization 0.30).



530