Adaptation of Host Transmission Cycle During Pathogen Speciation

Nitin Kumar[1,*,§], Hilary P. Browne[1,*], Elisa Viciani[1], Samuel C. Forster[1,2,3], Simon Clare[4], Katherine Harcourt[4], Mark D. Stares[1], Gordon Dougan[4], Derek J. Fairley[5], Paul Roberts[6], Munir Pirmohamed[6], Martha RJ Clokie[7], Mie Birgitte Frid Jensen[8], Katherine R. Hargreaves[7], Margaret Ip[9], Lothar H. Wieler[10,11], Christian Seyboldt[12], Torbjörn Norén[13,14], Thomas V. Riley[15,16], Ed J. Kuijper[17], Brendan W. Wren[18], Trevor D. Lawley[1,§]

[1]Host-Microbiota Interactions Laboratory, Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
[2]Centre for Innate Immunity and Infectious Diseases, Hudson Institute of Medical Research, Clayton, Victoria, 3168, Australia.
[3]Department of Molecular and Translational Sciences, Monash University, Clayton, Victoria, 3800, Australia.
[4]Wellcome Sanger Institute, Hinxton, CB10 1SA, UK.
[5]Belfast Health and Social Care Trust, Belfast, Northern Ireland.
[6]University of Liverpool, Liverpool, UK.
[7]Department of Infection, Immunity and Inflammation, University of Leicester, Leicester, LE1 7RH, UK.
[8]Department of Clinical Microbiology, Slagelse Hospital, Ingemannsvej 18, 4200, Slagelse, Denmark.
[9]Department of Microbiology, Chinese University of Hong Kong, Shatin, Hong Kong.
[10]Institute of Microbiology and Epizootics, Department of Veterinary Medicine, Freie Universität Berlin, Berlin, Germany.
[11]Robert Koch Institute, Berlin, Germany.
[12]Institute of Bacterial Infections and Zoonoses, Federal Research Institute for Animal Health (Friedrich-Loeffler-Institut), Jena, Germany.
[13]Faculty of Medicine and Health, Örebro University, Örebro, Sweden.
[14]Department of Laboratory Medicine, Örebro University Hospital Örebro, Sweden
[15]Department of Microbiology, PathWest Laboratory Medicine, Queen Elizabeth II Medical Centre, Western Australia, Australia.
[16]School of Pathology & Laboratory Medicine, The University of Western Australia, Western Australia, Australia
[17]Section Experimental Bacteriology, Department of Medical Microbiology, Leiden University Medical Center, Leiden, Netherlands.
[18]Department of Pathogen Molecular Biology, London School of Hygiene and Tropical Medicine, University of London, London, UK.

[*]These authors contributed equally to this work

[§]Corresponding authors

Trevor D. Lawley: Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK, CB10 1SA, Phone 01223 495 391, Fax 01223 495 239, Email: tl2@sanger.ac.uk

Nitin Kumar: Wellcome Sanger Institute, Hinxton, Cambridgeshire, UK, CB10 1SA, Phone 01223 495 391, Fax 01223 495 239, Email: nk6@sanger.ac.uk

Bacterial speciation is a fundamental evolutionary process characterized by diverging genotypic and phenotypic properties. However, the selective forces impacting the genetic adaptations and how they relate to the biological changes that underpin the formation of a new bacterial species remain poorly understood. Here we reveal that the spore-forming, healthcare-associated enteropathogen *Clostridium difficile* is actively undergoing speciation, and that diverging genetic lineages with distinct transmission properties formed prior to the advent of the modern healthcare system. Applying large-scale genomic analysis of 906 strains, we demonstrate that the ongoing speciation process is linked to positive selection on core genes in the newly forming species that are involved in spore formation and structure, and the metabolism of simple dietary sugars. Functional validation demonstrates the new *C. difficile* produce more resistant spores and show increased sporulation and host colonization capacity when glucose or fructose are available for metabolism. Thus, we reveal the formation of a new *C. difficil*e species, selected for metabolizing simple dietary sugars and producing high levels of resistant spores, that is specialized for healthcare-mediated transmission.

66 The formation of a new bacterial species from its ancestor is characterised by genetic
67 diversification and biological adaptation[1-4]. For decades, a polyphasic examination[5], relying on
68 genotypic and phenotypic properties of a bacterium, has been used to define and discriminate
69 a "species". The bacterial taxonomic classification framework has more recently used large
70 scale genome analysis to incorporate aspects of a bacterium's natural history, such as ecology[6],
71 horizontal gene transfer[1], recombination[2] and phylogeny[3]. Although a more accurate definition
72 of a bacterial species can be achieved with whole genome-based approaches, we still lack a
73 fundamental understanding of how selective forces impact adaptation of biological pathways
74 and phenotypic changes leading to bacterial speciation. In this work, we describe a unique
75 example of genome evolution and biological changes during the ongoing formation of a new
76 *C. difficile* species that is highly specialised for human transmission in the modern healthcare
77 system.

78 *C. difficile* is a strictly anaerobic, Gram-positive bacterial species that produces highly
79 resistant, metabolically dormant spores capable of rapid transmission between mammalian
80 hosts through environmental reservoirs[7]. Over the past four decades, *C. difficile* has emerged
81 as the leading cause of antibiotic-associated diarrhoea worldwide, with a large burden on the
82 healthcare system[7,8]. To define the evolutionary history and genetic changes underpinning the
83 emergence of *C. difficile* as a healthcare pathogen, we performed whole genome sequence
84 analysis of 906 strains isolated from humans (n=761), animals (n=116) and environmental
85 sources (n=29) with representatives from 33 countries and the largest proportion originating
86 from the UK (n=465) (Supplementary Fig. 1; Supplementary Table 1; Supplementary Table
87 2). This data is summarized visually here https://microreact.org/project/H1QidSp14. Our
88 collection was designed to capture comprehensive *C. difficile* genetic diversity[9] and includes
89 13 high-quality and well-annotated reference genomes (Supplementary Table 2). Robust
90 maximum likelihood phylogeny based on 1,322 concatenated single copy core genes (Fig. 1a;

91 Supplementary Table 3) illustrates the existence of four major phylogenetic groups within this
92 collection. Bayesian analysis of population structure (BAPS) using concatenated alignment of
93 1,322 single copy core genes corroborated the presence of the four distinct phylogenetic
94 groupings (PGs 1-4) (Fig. 1a) that each harbour strains from different geographical locations,
95 hosts and environmental sources which indicates signals of sympatric speciation. Each
96 phylogenetic group also harbours distinct clinically relevant ribotypes (RT): PG1 (RT001, 002,
97 014); PG2 (RT027 and 244); PG3 (RT023 and 017); PG4 (RT078, 045 and 033).

98 The phylogeny was rooted using closely related species (*C. bartlettii, C. hiranonis*, *C.*
99 *ghonii* and *C. sordellii*) as outgroups (Fig. 1a). This analysis indicated that three phylogenetic
100 groups (PG1, 2 and 3) of *C. difficile* descended from the most diverse phylogenetic group
101 (PG4). This was also supported by the frequency of SNP differences in pairwise comparisons
102 between strains of PG4 and each of the other PGs versus the level of pairwise SNP differences
103 between comparisons of PGs 1, 2 and 3 to each other (Supplementary Fig. 2). Interestingly,
104 bacteria from PG4 display distinct colony morphologies compared to bacteria from PG 1, 2
105 and 3 when grown on nutrient agar plates (Supplementary Fig. 3), suggesting a link between
106 *C. difficile* colony phenotype and genotype that distinguishes PG 1, 2 and 3 from PG4.

107 Our previous genomic study using 30 *C. difficile* genomes indicated an ancient,
108 genetically diverse species that likely emerged 1 to 85 million years ago[10]. Testing this estimate
109 using our larger dataset indicated the species emerged approximately 13.5 million years (12.7
110 - 14.3 million) ago. Using the same BEAST[11] analysis on our substantially expanded collection,
111 we estimate the most recent common ancestor (MRCA) of PG4 (using RT078 clade) arose
112 approximately 385,000 (297,137 - 582,886) years ago. In contrast, the MRCA of the PG1, 2
113 and 3 groups (using RT027 clade) arose approximately 76,000 (40,220 – 214,555) years ago.
114 Bayesian skyline analysis reveals a population expansion of PG1, 2 and 3 groups (using RT027
115 clade) around 1595 A.D., which occurred shortly before the emergence of the modern

116 healthcare system in the 18th century (Supplementary Fig. 4). Combined, these observations
117 suggest that PG4 emerged prior to the other PGs and that the PG1, 2 and 3 population structure
118 started to expand just prior to the implementation of the modern healthcare system[12]. We
119 therefore refer to PG4 as the "old" *C. difficile* and the PG1, 2 and 3 groups are referred to as
120 "new" *C. difficile.*

121 To investigate genomic relatedness, we next performed pairwise Average Nucleotide
122 Identity (ANI) analysis (Fig. 1b). This analysis revealed high nucleotide identity (ANI > 95%)
123 between PGs 1, 2 and 3 indicating that bacteria from these groups belong to the same species;
124 however, ANI between PG4 and any other PG was either less than the species threshold (ANI
125 > 95%) or on the borderline of the species threshold (94.04 - 96.25%) (Fig. 1b). To detect
126 recombination events within and between old and new *C. difficile,* FastGEAR analysis[13] was
127 performed on whole genome sequences of 906 strains (Supplementary Fig. 5). While analysis
128 identified increased recombination within new *C. difficile* lineages (PG1 - PG2: 1 - 102, PG1
129 - PG3: 1 - 214, PG2 - PG3: 1 - 96) (Supplementary Fig. 5) a restricted number of recombination
130 events between old and new *C. difficile* was observed (PG1 - PG4: 1 - 20, PG2 - PG4: 1 - 25,
131 PG3 - PG4: 1 - 46). This analysis strongly indicates the presence of recombination barriers in
132 the core genome that further distinguishes new *C. difficile* from old *C. difficile* and could
133 encourage sympatric speciation.

134 Functional analysis of the accessory genomes also shows a clear separation between
135 new and old *C. difficile* (Supplementary Fig. 6a). Cell motility (including flagella) and mobile
136 element functions are the most enriched functions in the accessory genome of new *C. difficile*
137 (Supplementary Fig. 6b; Supplementary Table 4), whereas the accessory genome of old *C.*
138 *difficile* is dominated by the uncharacterized function and DNA replication and modification
139 functions (Supplementary Fig. 6c; Supplementary Table 5). We also observe a higher number
140 of pseudogenes in new *C. difficile* compared to old *C. difficile* (Supplementary Fig. 7;

141 Supplementary Table 6-9). Comparative functional analysis of pseudogenes between old and
142 new *C. difficile* indicates phage-related function (n=13/24) is the largest functional category in
143 new *C. difficile* (Supplementary Table 10), whereas old *C. difficile* is dominated by
144 uncharacterized function (n=68/90) and transposons (13/90) (Supplementary Table 11). These
145 results indicate different selection pressures on the accessory genomes of new *C. difficile* from
146 old *C. difficile*.

147 In addition to reduced rates of recombination events, advantageous genetic variants in
148 a population driven by positive selective pressures, termed positive selection, are also a marker
149 of speciation[6]. We determined the Ka/Ks ratios and identified 172 core genes in new *C. difficile*
150 and 93 core genes in old *C. difficile* that were positively selected (Ka/Ks >1) (Fig. 2a;
151 Supplementary Table 12; Supplementary Table 13). Functional annotation and enrichment
152 analysis identified positively selected genes involved in carbohydrate and amino acid
153 metabolism, sugar phosphotransferase system (PTS) and spore coat architecture and spore
154 assembly in new *C. difficile* (Fig. 2b). In contrast, the sulphur relay system was the only
155 enriched functional category found among the positively selected genes from the old *C. difficile*
156 lineage. Notably, 26% (45 in total) of the positively selected genes in new *C. difficile* produce
157 proteins that are either directly involved in spore production, are present in the mature spore
158 proteome[14] or are regulated by Spo0A[15] or its sporulation-specific sigma factors[16] (Fig. 2c). In
159 contrast, no positively selected genes are directly involved in spore production in old *C.*
160 *difficile*; however, 22.5% (21 genes in total) are either present in the mature spore proteome or
161 are regulated by Spo0A or its sporulation specific sigma factors (Supplementary Fig. 8). The
162 lack of overlap between sporulation-associated positively selected genes in the two lineages
163 suggests a divergence of spore-mediated transmission functions. In addition, these results
164 suggest functions important for host-to-host transmission have evolved in new *C. difficile.*

165 We found 20 positively selected genes (Supplementary Table 12) in new *C. difficile*
166 whose products are components of the mature spore[14,15] and could contribute to environmental
167 survival. As an example, *sodA* (superoxide dismutase A), a gene associated with spore coat
168 assembly[17], has three-point mutations which are present in all new *C. difficile* genomes but
169 absent in old *C. difficile* genomes (Supplementary Fig. 9). Spores derived from diverse *C.*
170 *difficile* clades have a wide variation in resistance to microbiocidal free radicals from gas
171 plasma[18]. To investigate if the phenotypic resistance properties of spores from the new lineage
172 have evolved, we exposed spores from new and old *C. difficile* lineages to hydrogen peroxide,
173 a commonly used healthcare environmental disinfectant[17]. Spores derived from new *C. difficile*
174 were statistically significantly more resistant to 3% (P=0.0317) and 10% hydrogen peroxide
175 (P=0.0317) when compared to spores from old *C. difficile*, although there was no difference in
176 survival at 30% peroxide likely due to the overpowering bactericidal effect at this concentration
177 (P= 0.1667) (Fig. 3a).
178 The master regulator of *C. difficile* sporulation, Spo0A, is under positive selection in
179 new *C. difficile* only. Spo0A also controls other host colonization factors, such as flagella, and
180 carbohydrate metabolism, potentially serving to mediate cellular processes to direct energy to
181 spore production and host colonization to facilitate host-to-host transmission[15]. Interestingly,
182 the new *C. difficile* genomes contain genes under positive selection that are involved in fructose
183 metabolism (*fruABC* and *fruK*), glycolysis (*pgk* and *pyk*), sorbitol (CD630_24170) and ribulose
184 metabolism (*rep1*), and conversion of pyruvate to lactate (*ldh*). To further explore the link
185 between sporulation and carbohydrate metabolism in new *C. difficile*, we analysed positively
186 selected genes using KEGG pathways[19] and manual curation. Manual curation of key enriched
187 pathways across the 172 positively selected core genes in new *C. difficile* identified a complete
188 fructose-specific PTS pathway and identified four genes (30%, 4/13) involved in anaerobic
189 glycolysis during glucose metabolism (Supplementary Fig. 10). Other genes associated with

190 enriched PTS pathways include genes used for the cellular uptake and metabolism of mannitol,
191 cellobiose, glucitol/sorbitol, galactitol, mannose and ascorbate. Furthermore, comparative
192 analysis of carbohydrate active enzymes (CAZymes)[20] identified a clear separation of
193 CAZymes between new *C. difficile* and old *C. difficile* (Supplementary Fig. 11). Combined,
194 these observations suggest a divergence of functions between new and old *C. difficile* linked
195 to metabolism of a broad range of simple dietary sugars used in modern Western society[21].

196 The simple sugars glucose and fructose are increasingly used in diets within Western
197 societies[21]. Interestingly, trehalose, a disaccharide of glucose, used as a food additive has
198 impacted the emergence of some human virulent *C. difficile* variants[22]. Based on our genomic
199 analysis, we hypothesized that dietary glucose or fructose could differentially impact host
200 colonization by spores from new or old *C. difficile*. We therefore supplemented the drinking
201 water of mice with either glucose, fructose or ribose and challenged with new or old *C. difficile*
202 strains. Ribose metabolic genes were not under positive selection so this sugar was included as
203 a control. Mice challenged with new *C. difficile* spores exhibited statistically significant
204 increased bacterial load when exposed to dietary glucose (P= 0.048) or fructose (P= 0.0045)
205 compared to old *C. difficile* (Fig. 3b). No difference in bacterial load was observed between
206 new and old *C. difficile* without supplemented sugars or when supplemented with ribose (P=
207 0.2709) (Fig. 3b).

208 The infectivity and transmission of *C. difficile* within healthcare settings is facilitated
209 by environmental spore density[23,24]. To determine the impact of simple sugar availability on
210 spore production rates we assessed the ability of the two lineages to form spores in basal
211 defined medium (BDM) alone or supplemented with either glucose, fructose or ribose. While
212 no difference was observed on the ribose control (P= 0.3095), new *C. difficile* strains exhibited
213 statistically significant increased spore production on glucose (P= 0.0317) and fructose (P=
214 0.0317) (Fig. 3c). These results provide experimental validation and, together with our genomic

215 predictions, suggest that enhanced host colonization and onward spore-mediated transmission
216 with the consumption of simple dietary sugars is a feature of new *C. difficile* but not old *C.*
217 *difficile*.

218 The rapid recent emergence of *C. difficile* as a significant healthcare pathogen has
219 mainly been attributed to the genomic acquisition of antibiotic resistance and carbohydrate
220 metabolic functions on mobile elements via horizontal gene transfer[22,25]. Here we show that
221 these recent genomic adaptations have occurred in established, distinct evolutionary lineages
222 each with core genomes expressing unique, pre-existing transmission properties. We reveal the
223 ongoing formation of a new species, which we refer to as new *C. difficile*, with biological and
224 phenotypic properties consistent with a transmission cycle that was primed for human
225 transmission in the modern healthcare system (Fig. 3d). Indeed, different transmission
226 dynamics and host epidemiology have also been reported for new *C. difficile* (027 clade[26] and
227 017 clade[27]) endemic in healthcare systems in different parts of the world, and the 078 clade
228 that likely enters the human population from livestock[28-30]. Further, broad epidemiological
229 screens of *C. difficile* present in the healthcare system often highlight high abundances of new
230 *C. difficile* lineages as they represent 68.5% (USA), 74% (Europe) and 100% (Mainland China)
231 of the infecting strains[7,8,31,32]. Thus, we reveal a link between new *C. difficile* speciation,
232 adapted biological pathways and epidemiological patterns. In summary, our study elucidates
233 how bacterial speciation may prime lineages to emerge and transmit in a process accelerated
234 by modern human diet, the acquisition of antibiotic resistance or healthcare regimes.

235

236

237

238

239 **Materials and Methods:**

240 **Collection of *C. difficile* strains**

241 Laboratories worldwide were asked to send a diverse representation of their *C. difficile*
242 collections to the Wellcome Sanger Institute (WSI). After receiving all shipped samples the
243 DNA extraction was performed batch-wise using the same protocol and reagents to minimize
244 bias. Phenol-Chloroform was the preferred method for extraction since it provides high DNA
245 yield and intact chromosomal DNA.

246 The genomes of 382 strains designated as *C. difficile*, by PCR ribotyping were sequenced and
247 combined with our previous collection of 506 *C. difficile* strains, 13 high quality *C. difficile*
248 reference strains and 5 publicly available *C. difficile* RT 244 strains making a total of 906
249 strains analyzed in this study. This genome collection includes strains from humans (n=761),
250 animals (n=116) and the environment (n=29) that were collected from diverse geographic
251 locations (UK; n= 465, Europe; n= 230, N-America; n= 111, Australia; n= 62, Asia; n= 38).
252 Details of all strains are listed in Supplementary Table 1 and Supplementary Table 2, including
253 the European Nucleotide Archive (ENA) sample accession numbers. Metadata of this *C.*
254 *difficile* collection has been made freely publicly available through Microreact[33]
255 (https://microreact.org/project/H1QidSp14). The 13 *C. difficile* reference isolates
256 (Supplementary Table 2) are publicly available from the National Collection of Type Cultures
257 (NCTC) and the annotation of these genomes are available from the Host-Microbiota
258 Interactions Lab (HMIL; www.lawleylab.com), WSI.

259 **Bacterial culture and genomic DNA preparation**

260 *C. difficile* strains were cultured on blood agar plates for 48 hours, inoculated into
261 brain–heart infusion broth supplemented with yeast extract and cysteine and grown overnight
262 (16 hours) anaerobically at 37 °C. Cells were pelleted, washed with PBS, and genomic DNA
263 preparation was performed using a phenol–chloroform extraction as previously described[34].

264 All culturing of *C. difficile* took place in anaerobic conditions (10% CO_2, 10% H_2, 80% N_2)
265 in a Whitley DG250 workstation at 37 °C. All reagents and media were reduced for 24 hours
266 in anaerobic conditions before use.

267 **DNA sequencing, assembly and annotation**

268 Paired-end multiplex libraries were prepared and sequenced using Illumina Hi-Seq
269 platform with fragment size of 200-300 bp and a read length of 100 bp, as previously
270 described[35,36]. An in-house pipeline developed at the WTSI (https://github.com/sanger-
271 pathogens/Bio-AutomatedAnnotation) was used for bacterial assembly and annotation. It
272 consisted of *de novo* assembly for each sequenced genome using Velvet v1.2.10[37], SSPACE
273 v2.0[38] and GapFiller v1.1[39] followed by annotation using Prokka v1.5-1[40]. For the 13 high-
274 quality reference genomes, strains Liv024, TL178, TL176, TL174, CD305 and Liv022 were
275 sequenced using 454 and Illumina sequencing platforms, BI-9 and M68 were sequenced using
276 454 and capillary sequencing technologies with the assembled data for these 8 strains been
277 improved to an 'Improved High Quality Draft' genome standard[41]. Optical maps using the
278 Argus Optical Mapping system were also generated for Liv024, TL178, TL176, TL174, CD305
279 and Liv022. The remaining strains are all contiguous and were all sequenced using 454 and
280 capillary sequencing technologies except for R20291 which also had Illumina data
281 incorporated and 630 which was sequencing using capillary sequence data alone.

282 **Phylogenetic analysis, Pairwise SNP distances analysis and Average Nucleotide Identity**
283 **analysis**

284 The phylogenetic analysis was conducted by extracting nucleotide sequence of 1,322
285 single copy core gene from each *C. difficile* genome using Roary[42]. The nucleotide sequences
286 were concatenated and aligned with MAFFT v7.20[43]. Gubbins[44] was used to mask
287 recombination from concatenated alignment of these core genes and a maximum-likelihood
288 tree was constructed using RAxML v8.2.8[45] with the best-fit model of nucleotide substitution

289 (GTRGAMMA) calculated from ModelTest embedded in TOPALi v2.5[46] and 500 bootstrap
290 replicates. The phylogeny was rooted using a distance-based tree generated using Mash v2.0[47],
291 R package APE[48] and genome assemblies of closely related species (*C. bartlettii, C. hiranonis*,
292 *C. ghonii* and *C. sordellii*). All phylogenetic trees were visualized in iTOL[49]. Genomes of
293 closely related *C. difficile* were downloaded from NCBI. Pairwise SNP distances analysis was
294 performed on concatenated alignment of 1,322 single copy core genes using SNP-Dist
295 (https://github.com/tseemann/snp-dists). Average nucleotide analysis (ANI) was calculated by
296 performing pairwise comparison of genome assemblies using MUMmer[50].

297 **Population structure and recombination analysis**

298 Population structure based on concatenated alignment of 1,322 single copy core genes
299 of *C. difficile* was inferred using the HierBAPS[51] with one clustering layers and 5, 10 and 20
300 expected numbers of clusters (k) as input parameters. Recombination events across the whole
301 genome sequences were detected by mapping genomes against a reference genome (NCTC
302 13366; RT027) and using FastGear[13] with default parameters.

303 **Functional genomic analysis**

304 To explore accessory genome and identify protein domains in a genome, we performed
305 RPS-BLAST using COG database (accessed February 2019)[52]. All protein domains were
306 classified in different functional categories using the COG database[52] and were used to perform
307 Discriminant Analysis of Principle Components (DAPC)[53] implemented in the R package
308 Adegenet v2.0.1[54]. Domain and functional enrichment analysis was calculated using one-sided
309 Fisher's exact test with p-value adjusted by Hochberg method in R v3.2.2.

310 Carbohydrate active enzymes (CAZymes) in a genome were identified using dbCAN
311 v5.0[55] (HMM database of carbohydrate active enzyme annotation). Best hits include hits with
312 E-value < 1e-5 if alignment > 80 aa and hits with E-value < 1e-3 if alignment < 80 aa, and

313 alignment coverage > 0.3. Best hits were used to perform Discriminant Analysis of Principle
314 Components (DAPC)[53] implemented in the R package Adegenet v2.0.1[54].
315 Functional annotation of positively selected genes was carried out using the Riley
316 classification system[56], KEGG Orthology[57] and Pfam functional families[58].
317 **Analysis of selective pressures.**
318 The aligned nucleotide sequences of each 1,322 single copy core genes were extracted
319 from Roary's output. The ratio between the number of non-synonymous mutations (Ka) and
320 the number of synonymous mutations (Ks) was calculated for the whole alignment and for the
321 respective subsets of strains belonging to the PG1, 2, 3 as a group and PG4. The Ka/Ks ratio
322 for each gene alignment was calculated with SeqinR v3.1. A Ka/Ks > 1 was considered as the
323 threshold for identifying genes under positive selection.
324 **Pseudogenes analysis**
325 Nucleotide annotations of genes within a genome within each phylogenetic group were
326 mapped against the protein sequences of the reference genome for its phylogenetic group (PG1:
327 NCTC 13307(RT012), PG2: SRR2751302 (RT244), PG3: NCTC 14169 (RT017), PG4: NCTC
328 14173 (RT078)) using TBLASTN as previously described[59]. Pseudogenes were called based
329 on following criteria: genes with E value > 1-30 and sequence identity < 99% and which are
330 absent in 90% members of a phylogenetics group. Genes in the reference genomes annotated
331 as a pseudogene were also included in addition to genes in query genomes.
332 **Analysis of estimating dates**
333 The aligned nucleotide sequences of each 222 core genes of *C. difficile* which are under
334 neutral selection (Ka/Ks = 1) were extracted from Roary's output. Gubbins[44] was used to mask
335 recombination from concatenated alignment of these core genes and used as an input for
336 Bayesian Evolutionary Analysis Sampling Trees (BEAST) software package v2.4.1[11]. In
337 BEAST, the MCMC chain was run for 50 million generations, sampling every 1000 states

338 using the strict clock model (2.50×10^{-9} - 1.50×10^{-8} per site per year)[10] and HKY four discrete
339 gamma substitution model, each run in triplicate. Convergence of parameters were verified
340 with Tracer v1.5[60] by inspecting the Effective Sample Sizes (ESS > 200). LogCombiner was
341 used to remove 10% of the MCMC steps discarded as burn-ins and combine triplicates. The
342 resulting file was used to infer the time of divergence from the most recent common ancestor
343 for *C. difficile*, old and new *C. difficile*. The Bayesian skyline plot was generated with Tracer
344 v1.5[60].

345 ***C. difficile* growth *in vitro* on selected carbon sources**

346 Basal defined medium (BDM)[61] was used as the minimal medium to which selected
347 carbon sources (2 g/L of glucose, fructose or ribose from Sigma-Aldrich) were added. *C.*
348 *difficile* strains were grown on CCEY agar (Bioconnections) for two days; 125-ml Erlenmeyer
349 flasks containing 10 mL of BDM with or without carbon sources were inoculated with *C.*
350 *difficile* strains and incubated in anaerobic conditions at 37º C shaking at 180 rpm. After 48
351 hours, spores were enumerated by centrifuging the culture to a pellet, carefully decanting the
352 BDM and re-suspending in 70% ethanol for 4 hours to kill vegetative cells. Following ethanol
353 shock, spores were washed twice in PBS and plated in a serial dilution on YCFA media
354 supplemented with 0.1% sodium taurocholate. Colony forming units (representing germinated
355 spores) were counted 24 hours later. Experiment was performed using 3 biological replicates
356 for each strain. New strains used were TL178 (RT002/ PG1), TL174 (RT015/ PG1), R20291
357 (RT027/ PG2), CF5 (RT017/ PG3) and CD305 (RT023/ PG3). Old strains used were MON024
358 (RT033), CDM120 (RT078), WA12 (RT291), WA13 (RT228) and MON013 (RT127), all
359 PG4. Data was presented using GraphPad Prism v7.03.

360 ***C. difficile* spore resistance to disinfectant**

361 Spores were prepared by adapting the previous protocol[18]. In brief, *C. difficile* strains
362 were streaked on CCEY media, the cells were harvested from the plates 48 hours later and

363 subjecting to exposure in 70% ethanol for 4 hours to kill vegetative cells. The solution was
364 then centrifuged, ethanol was decanted and the spores were washed once in 5ml sterile saline
365 (0.9% w/v) solution before being suspended in 5ml of saline (0.9% w/v) with Tween20 (0.05%
366 v/v). 300ul spore suspensions (at a concentration of approximately 10^6 spores) were exposed
367 to 300ul of 3%, 10% and 30% hydrogen peroxide (Fisher Scientific UK Limited) solutions for
368 5 minutes in addition to 300ul PBS. The suspensions were then centrifuged, hydrogen peroxide
369 or PBS was decanted and the spores were washed twice with PBS. Washed spores were plated
370 on YCFA media with 0.1% sodium taurocholate to stimulate spore germination and colony
371 forming units were counted 24 hours later. Experiment was performed using 3 biological
372 replicates for each strain. New strains used were TL178 (RT002/ PG1), TL174 (RT015/ PG1),
373 R20291 (RT027/ PG2), CF5 (RT017/ PG3) and CD305 (RT023/ PG3). Old strains used were
374 MON024 (RT033), CDM120 (RT078), WA12 (RT291), WA13 (RT228) and MON013
375 (RT127), all PG4. Data was presented using GraphPad Prism v7.03.

376 ***In vivo C. difficile* colonisation experiment**

377 Five female 8-week-old C57BL/6 mice were given 250 mg/L clindamycin (Apollo
378 Scientific) in drinking water. After 5 days, clindamycin treatment was interrupted and 100 mM
379 of glucose, fructose or ribose was added to mouse drinking water for the rest of the experiment;
380 no sugars were given to control mice. After 3 days, mice were infected orally with 6 x 10^3
381 spore/mouse of *C. difficile* R20291 (RT027) or M120 (RT078) strain. Faecal samples were
382 collected from all mice before infection to check for pre-existing *C. difficile* contamination.
383 Spore suspensions were prepared as described above[18]. After 16 hours, faecal samples were
384 collected from all mice to determine viable *C. difficile* cell counts by serial dilution and plating
385 on CCEY agar supplemented with 0.1% sodium taurocholate. Data was presented using
386 GraphPad Prism version 7.03. Mouse experiments were approved by the Wellcome Sanger
387 Institute.

388 **References:**

389 1. Lawrence, J.G. & Retchless, A.C. The interplay of homologous recombination and
390 horizontal gene transfer in bacterial speciation. *Methods Mol Biol* **532**, 29-53 (2009).
391 2. Fraser, C., Alm, E.J., Polz, M.F., Spratt, B.G. & Hanage, W.P. The bacterial species
392 challenge: making sense of genetic and ecological diversity. *Science* **323**, 741-6 (2009).
393 3. Staley, J.T. The bacterial species dilemma and the genomic-phylogenetic species
394 concept. *Philos Trans R Soc Lond B Biol Sci* **361**, 1899-909 (2006).
395 4. Moeller, A.H. *et al.* Cospeciation of gut microbiota with hominids. *Science* **353**, 380-
396 382 (2016).
397 5. Vandamme, P. *et al.* Polyphasic taxonomy, a consensus approach to bacterial
398 systematics. *Microbiol Rev* **60**, 407-38 (1996).
399 6. Cohan, F.M. & Perry, E.B. A systematics for discovering the fundamental units of
400 bacterial diversity. *Curr Biol* **17**, R373-86 (2007).
401 7. Martin, J.S., Monaghan, T.M. & Wilcox, M.H. Clostridium difficile infection:
402 epidemiology, diagnosis and understanding transmission. *Nat Rev Gastroenterol*
403 *Hepatol* **13**, 206-16 (2016).
404 8. Lessa, F.C., Winston, L.G., McDonald, L.C. & Emerging Infections Program, C.d.S.T.
405 Burden of Clostridium difficile infection in the United States. *N Engl J Med* **372**, 2369-
406 70 (2015).
407 9. Stabler, R.A. *et al.* Macro and micro diversity of Clostridium difficile isolates from
408 diverse sources and geographical locations. *PLoS One* **7**, e31559 (2012).
409 10. He, M. *et al.* Evolutionary dynamics of Clostridium difficile over short and long time
410 scales. *Proc Natl Acad Sci U S A* **107**, 7527-32 (2010).
411 11. Drummond, A.J., Suchard, M.A., Xie, D. & Rambaut, A. Bayesian phylogenetics with
412 BEAUti and the BEAST 1.7. *Mol Biol Evol* **29**, 1969-73 (2012).
413 12. Jackson, M. & Spray, E.C. Health and Medicine in the Enlightenment. (Oxford
414 University Press, 2012).
415 13. Mostowy, R. *et al.* Efficient Inference of Recent and Ancestral Recombination within
416 Bacterial Populations. *Mol Biol Evol* **34**, 1167-1182 (2017).
417 14. Lawley, T.D. *et al.* Proteomic and genomic characterization of highly infectious
418 Clostridium difficile 630 spores. *J Bacteriol* **191**, 5377-86 (2009).
419 15. Pettit, L.J. *et al.* Functional genomics reveals that Clostridium difficile Spo0A
420 coordinates sporulation, virulence and metabolism. *BMC Genomics* **15**, 160 (2014).
421 16. Fimlaid, K.A. *et al.* Global analysis of the sporulation pathway of Clostridium difficile.
422 *PLoS Genet* **9**, e1003660 (2013).
423 17. Lawley, T.D. *et al.* Use of purified Clostridium difficile spores to facilitate evaluation of
424 health care disinfection regimens. *Appl Environ Microbiol* **76**, 6895-900 (2010).
425 18. Connor, M. *et al.* Evolutionary clade affects resistance of Clostridium difficile spores
426 to Cold Atmospheric Plasma. *Sci Rep* **7**, 41814 (2017).
427 19. Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M. & Tanabe, M. KEGG as a
428 reference resource for gene and protein annotation. *Nucleic Acids Res* **44**, D457-62
429 (2016).
430 20. Cantarel, B.L. *et al.* The Carbohydrate-Active EnZymes database (CAZy): an expert
431 resource for Glycogenomics. *Nucleic Acids Res* **37**, D233-8 (2009).
432 21. Lustig, R.H., Schmidt, L.A. & Brindis, C.D. Public health: The toxic truth about sugar.
433 *Nature* **482**, 27-9 (2012).

434 22. Collins, J. *et al.* Dietary trehalose enhances virulence of epidemic Clostridium difficile.
435 *Nature* (2018).
436 23. Browne, H.P. *et al.* Culturing of 'unculturable' human microbiota reveals novel taxa
437 and extensive sporulation. *Nature* **533**, 543-546 (2016).
438 24. Merrigan, M. *et al.* Human hypervirulent Clostridium difficile strains exhibit increased
439 sporulation as well as robust toxin production. *J Bacteriol* **192**, 4904-11 (2010).
440 25. Sebaihia, M. *et al.* The multidrug-resistant human pathogen Clostridium difficile has a
441 highly mobile, mosaic genome. *Nat Genet* **38**, 779-86 (2006).
442 26. He, M. *et al.* Emergence and global spread of epidemic healthcare-associated
443 Clostridium difficile. *Nat Genet* **45**, 109-13 (2013).
444 27. Cairns, M.D. *et al.* Comparative Genome Analysis and Global Phylogeny of the Toxin
445 Variant Clostridium difficile PCR Ribotype 017 Reveals the Evolution of Two
446 Independent Sublineages. *J Clin Microbiol* **55**, 865-876 (2017).
447 28. Dingle, K.E. *et al.* A Role for Tetracycline Selection in Recent Evolution of Agriculture-
448 Associated Clostridium difficile PCR Ribotype 078. *MBio* **10**(2019).
449 29. Knetsch, C.W. *et al.* Zoonotic Transfer of Clostridium difficile Harboring Antimicrobial
450 Resistance between Farm Animals and Humans. *J Clin Microbiol* **56**(2018).
451 30. Knight, D.R., Squire, M.M. & Riley, T.V. Nationwide surveillance study of Clostridium
452 difficile in Australian neonatal pigs shows high prevalence and heterogeneity of PCR
453 ribotypes. *Appl Environ Microbiol* **81**, 119-23 (2015).
454 31. Bauer, M.P. *et al.* Clostridium difficile infection in Europe: a hospital-based survey.
455 *Lancet* **377**, 63-73 (2011).
456 32. Tang, C. *et al.* The incidence and drug resistance of Clostridium difficile infection in
457 Mainland China: a systematic review and meta-analysis. *Sci Rep* **6**, 37865 (2016).
458 33. Argimon, S. *et al.* Microreact: visualizing and sharing data for genomic epidemiology
459 and phylogeography. *Microb Genom* **2**, e000093 (2016).
460 34. Croucher, N.J. *et al.* Rapid pneumococcal evolution in response to clinical
461 interventions. *Science* **331**, 430-4 (2011).
462 35. Harris, S.R. *et al.* Evolution of MRSA during hospital transmission and intercontinental
463 spread. *Science* **327**, 469-74 (2010).
464 36. Quail, M.A. *et al.* A large genome center's improvements to the Illumina sequencing
465 system. *Nat Methods* **5**, 1005-10 (2008).
466 37. Zerbino, D.R. & Birney, E. Velvet: algorithms for de novo short read assembly using de
467 Bruijn graphs. *Genome Res* **18**, 821-9 (2008).
468 38. Boetzer, M., Henkel, C.V., Jansen, H.J., Butler, D. & Pirovano, W. Scaffolding pre-
469 assembled contigs using SSPACE. *Bioinformatics* **27**, 578-9 (2011).
470 39. Boetzer, M. & Pirovano, W. Toward almost closed genomes with GapFiller. *Genome*
471 *Biol* **13**, R56 (2012).
472 40. Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-9
473 (2014).
474 41. Chain, P.S. *et al.* Genomics. Genome project standards in a new era of sequencing.
475 *Science* **326**, 236-7 (2009).
476 42. Page, A.J. *et al.* Roary: rapid large-scale prokaryote pan genome analysis.
477 *Bioinformatics* **31**, 3691-3 (2015).
478 43. Katoh, K. & Standley, D.M. MAFFT multiple sequence alignment software version 7:
479 improvements in performance and usability. *Mol Biol Evol* **30**, 772-80 (2013).

44. Croucher, N.J. *et al.* Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* **43**, e15 (2015).
45. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312-3 (2014).
46. Milne, I. *et al.* TOPALi v2: a rich graphical interface for evolutionary analyses of multiple alignments on HPC clusters and multi-core desktops. *Bioinformatics* **25**, 126-7 (2009).
47. Ondov, B.D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132 (2016).
48. Popescu, A.A., Huber, K.T. & Paradis, E. ape 3.0: New tools for distance-based phylogenetics and evolutionary analysis in R. *Bioinformatics* **28**, 1536-7 (2012).
49. Letunic, I. & Bork, P. Interactive Tree Of Life v2: online annotation and display of phylogenetic trees made easy. *Nucleic Acids Res* **39**, W475-8 (2011).
50. Delcher, A.L., Phillippy, A., Carlton, J. & Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Res* **30**, 2478-83 (2002).
51. Cheng, L., Connor, T.R., Siren, J., Aanensen, D.M. & Corander, J. Hierarchical and spatially explicit clustering of DNA sequences with BAPS software. *Mol Biol Evol* **30**, 1224-8 (2013).
52. Tatusov, R.L., Galperin, M.Y., Natale, D.A. & Koonin, E.V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res* **28**, 33-6 (2000).
53. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet* **11**, 94 (2010).
54. Jombart, T. adegenet: a R package for the multivariate analysis of genetic markers. *Bioinformatics* **24**, 1403-5 (2008).
55. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res* **40**, W445-51 (2012).
56. Riley, M. Functions of the gene products of Escherichia coli. *Microbiol Rev* **57**, 862-952 (1993).
57. Kanehisa, M., Sato, Y. & Morishima, K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* **428**, 726-731 (2016).
58. Finn, R.D. *et al.* Pfam: the protein families database. *Nucleic Acids Res* **42**, D222-30 (2014).
59. Lerat, E. & Ochman, H. Recognizing the pseudogenes in bacterial genomes. *Nucleic Acids Res* **33**, 3125-32 (2005).
60. Rambaut, A., Drummond, A.J., Xie, D., Baele, G. & Suchard, M.A. Posterior Summarization in Bayesian Phylogenetics Using Tracer 1.7. *Syst Biol* **67**, 901-904 (2018).
61. Karasawa, T., Ikoma, S., Yamakawa, K. & Nakamura, S. A defined growth medium for Clostridium difficile. *Microbiology* **141 (Pt 2)**, 371-5 (1995).

Acknowledgements:

This work was supported by the Wellcome Trust [098051]; the United Kingdom Medical Research Council [PF451 and MR/K000511/1] and the Australian National Health and Medical Research Council [1091097 to SF] and the Victorian government. The authors thank Scott Weese, Fabio Miyajima, Glen Songer, Thomas Louie, Julian Rood, and Nicholas M. Brown for *C. difficile* strains. The authors thank Anne Neville, Daniel Knight and Bastian Hornung for critical reading and comments. The authors would also like to acknowledge the support of the Wellcome Sanger Institute Pathogen Informatics Team.

Funding for open access charge: Wellcome Sanger Institute.

Author Contributions:

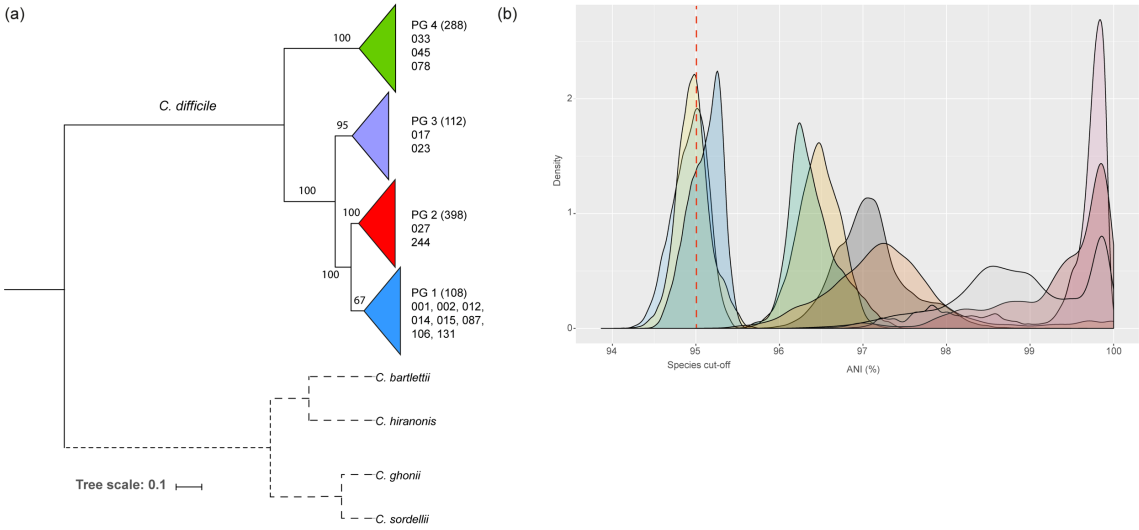
NK and TDL conceived and managed the study. NK, SCF, EV, HPB and TDL wrote the manuscript with input from other co-authors. NK performed the computational analysis. HPB performed genome annotation of reference genomes. EV, HPB, SCF and TDL designed *in vitro* and *in vivo* experiments. HPB, EV and MS performed *in vitro* experiments. EV, MDS, SC and KH performed *in vivo* experiments.

Conflict of interests

The authors declare no competing financial interests.

554 **Figures:**

555



556

557 **Figure 1. Phylogeny and population structure of *Clostridium difficile.*** (A) Maximum
558 likelihood tree of 906 *C. difficile* isolates constructed from the core genome alignment,
559 excluding recombination events. Collapsed clades as triangles represent four Phylogenetic
560 groups (PG1-4) identified by Bayesian analysis of population structure (BAPS). Number in
561 parentheses indicate number of isolates. Key PCR ribotypes in each PG are shown. Bootstrap
562 values of key branches are shown next to the branches. *Clostridium bartlettii*, *Clostridium*
563 *hiranonis, Clostridium ghonii* and *Clostridium sordellii* were used as outgroups to root the tree.
564 Scale bar indicates number of substitutions per site. (B) Distribution pattern of average
565 nucleotide identity (ANI) for 906 *C. difficile* isolates. Pairwise ANI calculations between
566 different PGs are shown in dark grey (PG1 and PG2), orange (PG1 and PG3), light blue (PG1
567 and PG4), light green (PG2 and PG3), light yellow (PG2 and PG4), cyan (PG3 and PG4).
568 Pairwise ANI calculations between strains of same PG are shown in dark orange (PG1), light
569 pink (PG2), light red (PG3) and light grey (PG4). Dotted red line indicates bacterial species
570 cut-off.

571

572

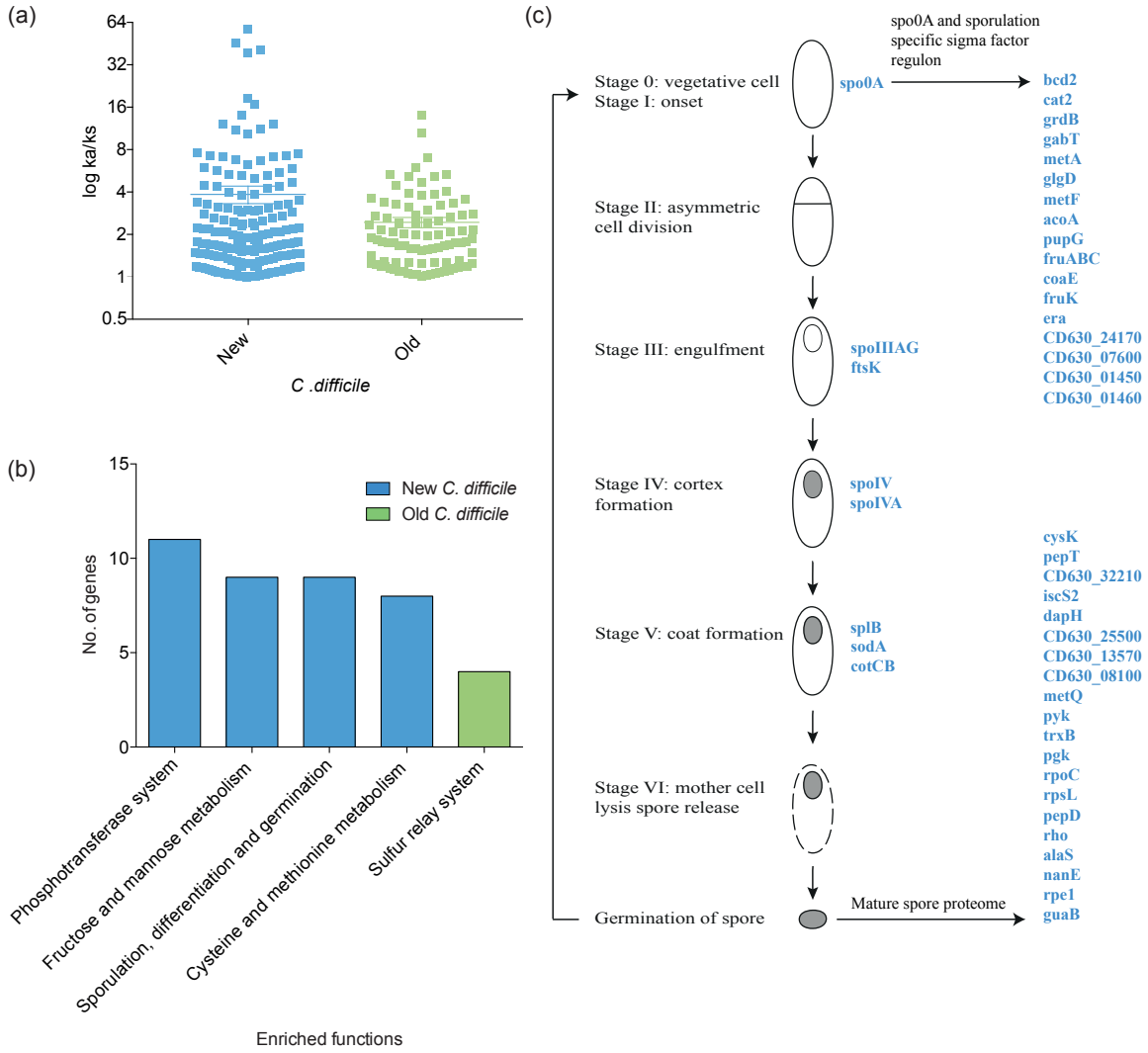


Figure 2. Adaptation of sporulation and metabolic genes in new *Clostridium difficile* lineage. Positive selection analysis of new and old *C. difficile* based on 1,322 core genes. (A) Distribution of Ka/Ks ratio for the positively selected genes in new *C. difficile* (n=172 genes) and old *C. difficile* (n=93 genes) is shown. Error bars are SEM. (B) Enriched functions in the positively selected genes of new (blue) and old (green) *C. difficile* are shown. Y–axis represent number of positive selected genes in each enriched function. All are statistically significant (sugar phosphotransferase system (q value < 1.7×10^{-3}), fructose and mannose metabolism (q value < 1.18×10^{-3}), sporulation, differentiation and germination (q value < 1.66×10^{-2}), cysteine and methionine metabolism (q value < 2.80×10^{-3}), sulphur relay system (q value < 8.00×10^{-3})). (C) Positively selected sporulation-associated genes in new *C. difficile* are shown

584 in blue. Of the 172 genes under positive selection, 26% (45 in total) are either involved in spore
585 production (sporulation stages I, III, IV and V), their proteins are present in the mature spore
586 proteome or they are regulated by Spo0A or its sporulation specific sigma factors.
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608

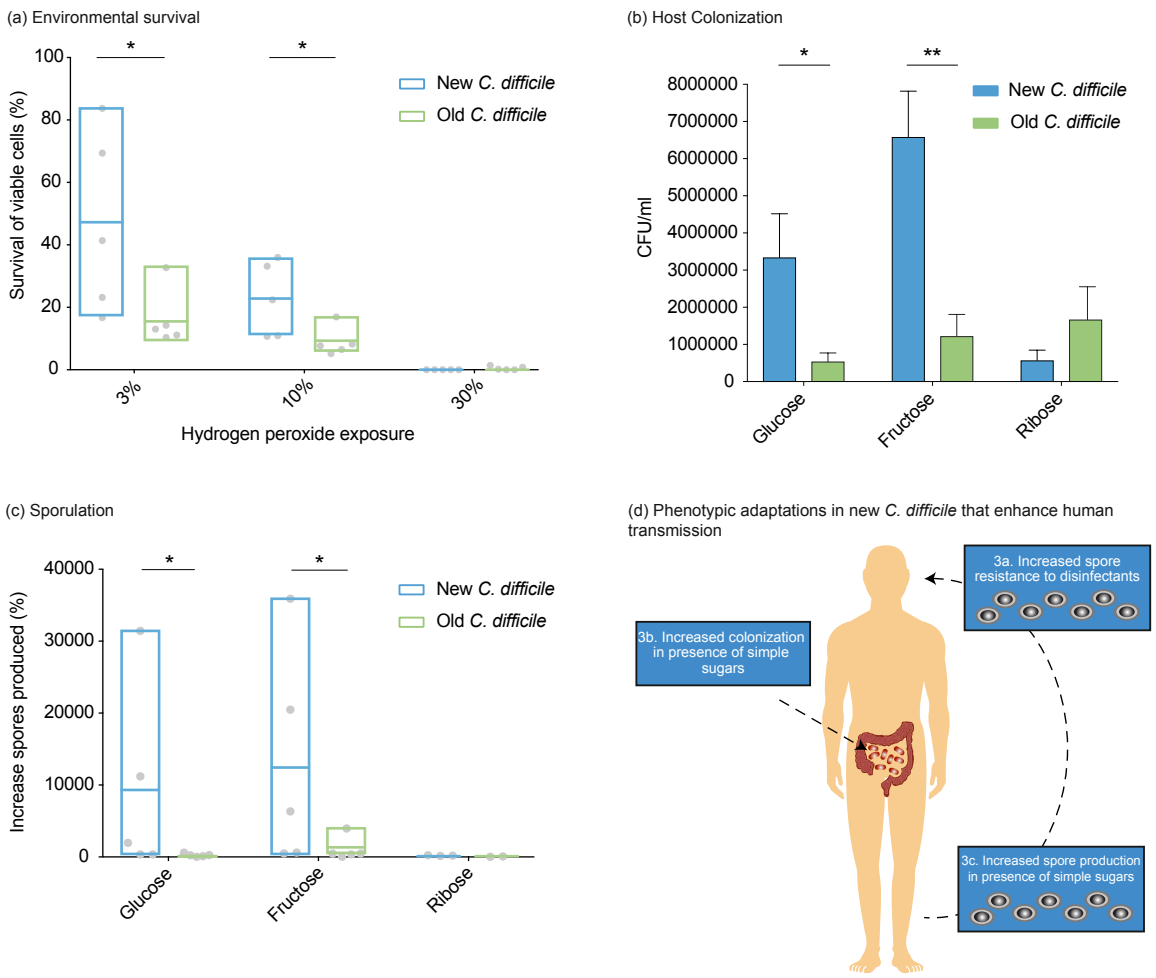
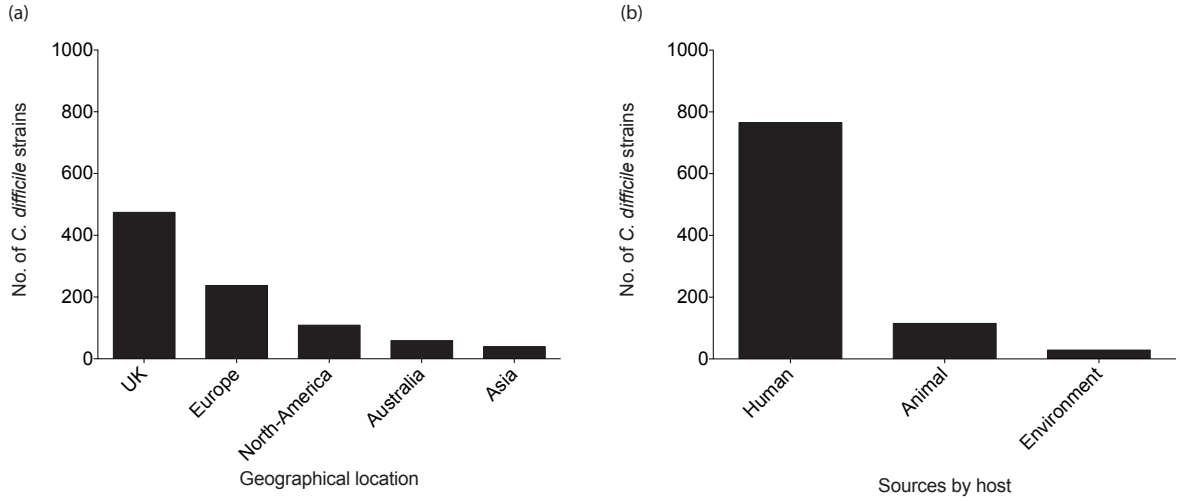


Figure 3. Bacterial speciation is linked with increased host adaptation and transmission ability. (A) Spores of new *C. difficile* are more resistant to widely used hydrogen peroxide disinfectant. Spores of new and old *C. difficile* (n=5 different ribotypes for both lineages) were exposed to hydrogen peroxide for 5 minutes, washed and plated. Recovered CFUs representing surviving germinated spores were counted and presented as a percentage of spores exposed to PBS. Mean and range, Mann-Whitney unpaired two-tailed test (* $P < 0.05$). **(B)** Intestinal colonisation of new strains is increased in the presence of simple sugars compared to old strains. Comparison of vegetative cell loads between new (n=1, RT027) and old (n=1, RT078) *C. difficile* strains in mice whose diet was supplemented with different sugars. CFUs from faecal samples cultured 16 hours after *C. difficile* challenge are presented. Mean values of 5 mice are shown, SEM, unpaired two-tailed t-test (* $P < 0.05$, **$P < 0.005$). **(C)** New strains

621 produce more spores in the presence of simple sugars. *C. difficile* new and old (n=5 different
622 ribotypes for both lineages) strains were grown on basal defined media in the presence or
623 absence of different sugars, vegetative cells were killed by ethanol exposure and the number
624 of CFUs representing germinated spores were counted. The percentage of spores recovered in
625 the presence of sugars compared to BDM alone is presented. Mean and range, Mann-Whitney
626 unpaired two-tailed test (*$P < 0.05$). (D) Overview of adaptations in key aspects of the new *C.*
627 *difficile* transmission cycle in human population.
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645

646 **Supplementary Figures:**

647



648

649 **Supplementary Figure 1. Breakdown of 906 *Clostridium difficile* isolates based on**

650 **metadata.** A. Number of strains based on geographical location is shown in bar-plots. B.

651 Number of strains based on source.

652

653

654

655

656

657

658

659

660

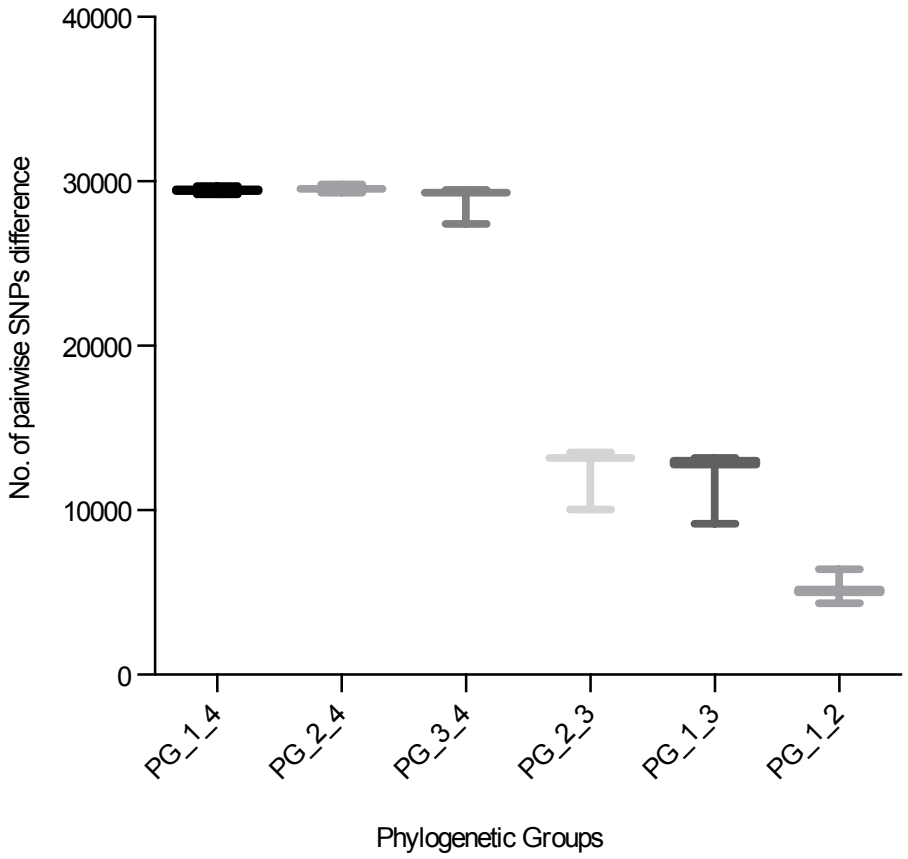
661

662

663

664

665

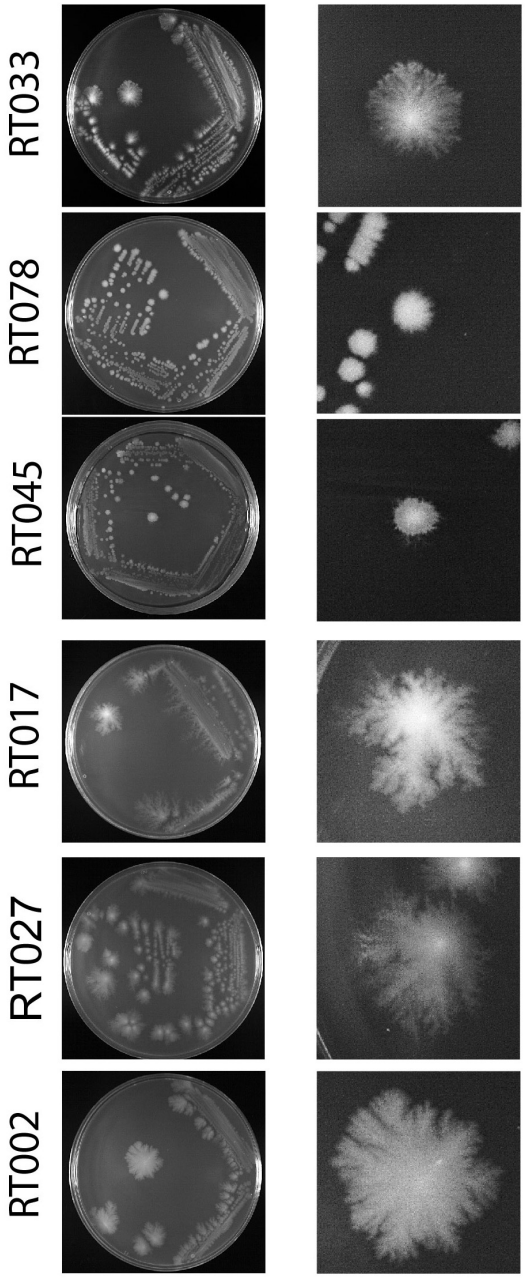


666

667 **Supplementary Figure 2. Pairwise SNPs difference between different phylogenetic**

668 **groups of *Clostridium difficile.*** Boxplots show distribution of SNPs differences calculated

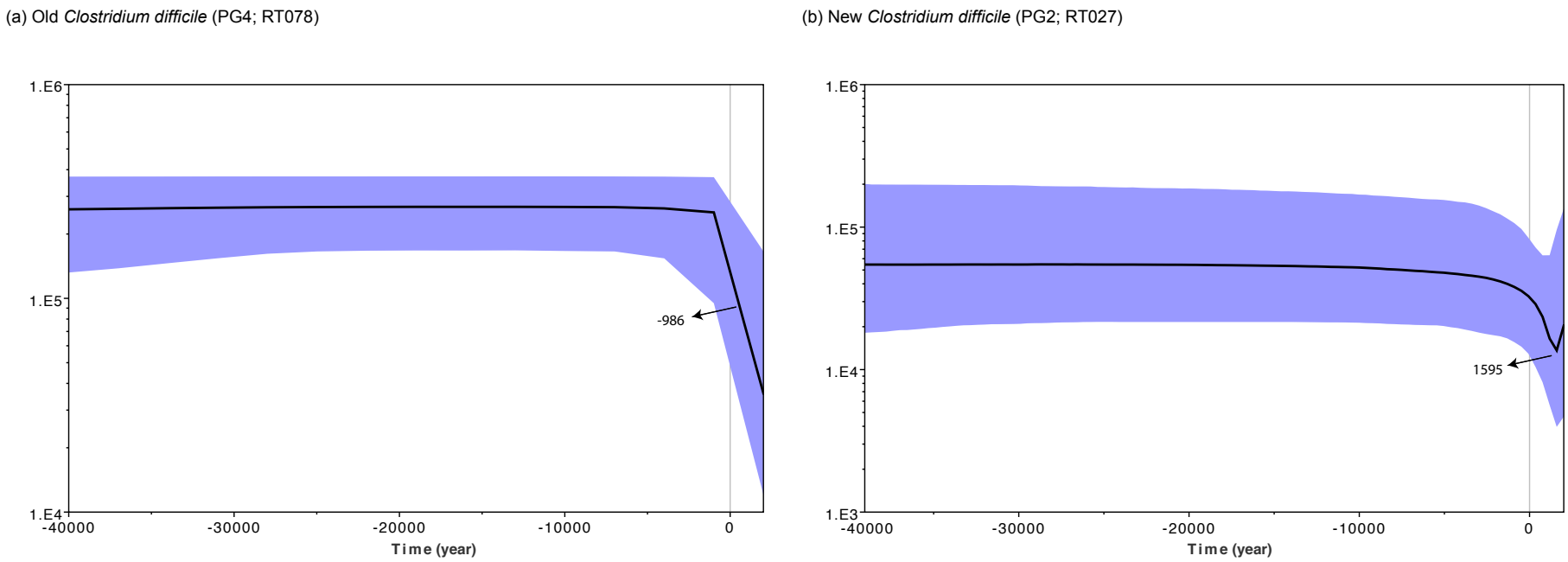
669 between pairs of genomes belonging to different PGs.



670

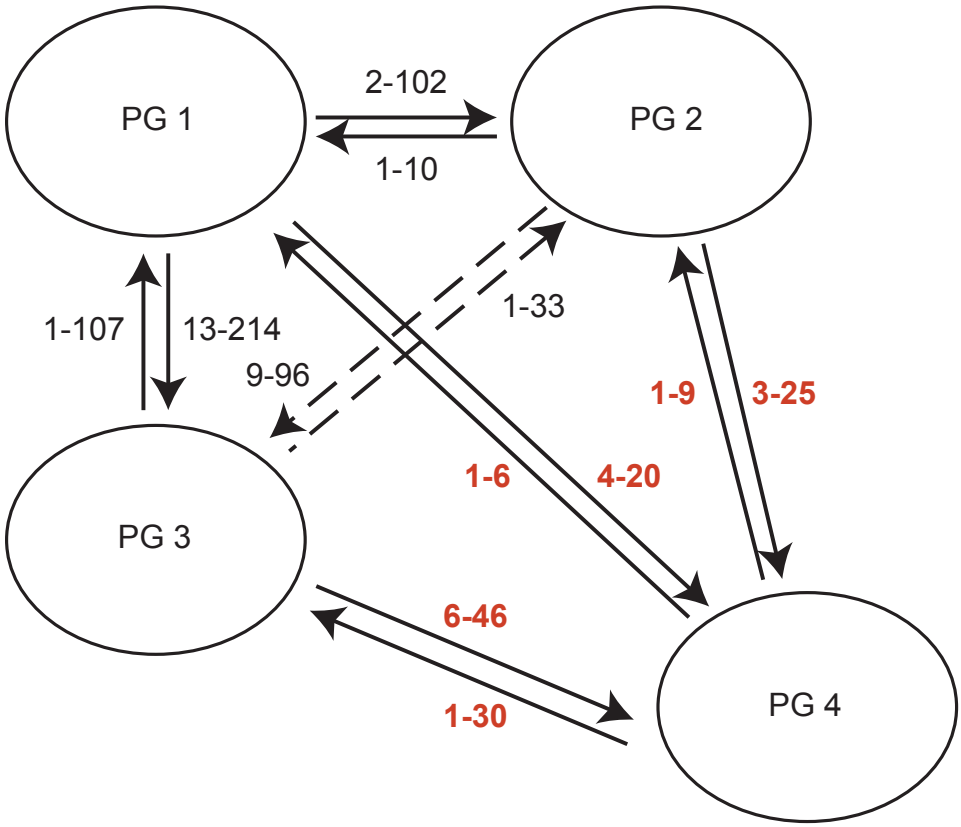
671 **Supplementary Figure 3. Colony morphology of *Clostridium difficile* strains.** *C. difficile*
672 strains from distinct clades were plated on YCFA agar plates supplemented with 0.1% sodium
673 taurocholate and incubated for 8 days and *C. difficile* colonies were photographed. Ribotype
674 (RT) 002, RT027, and RT017 represent PG1, 2 and 3 respectively. RT045, RT078 and RT033
675 represent PG4.

676



Supplementary Figure 4. Bayesian skyline plot of old (PG4; RT078) and new (PG2; RT027) *Clostridium difficile* indicate signals of new *C. difficile* expansion in the year 1595. The black line represents median estimate, and purple area represents its 95% highest posterior density intervals.

683



684

685 **Supplementary Figure 5.** Recombination analysis based on whole genome of 906 *Clostridium*

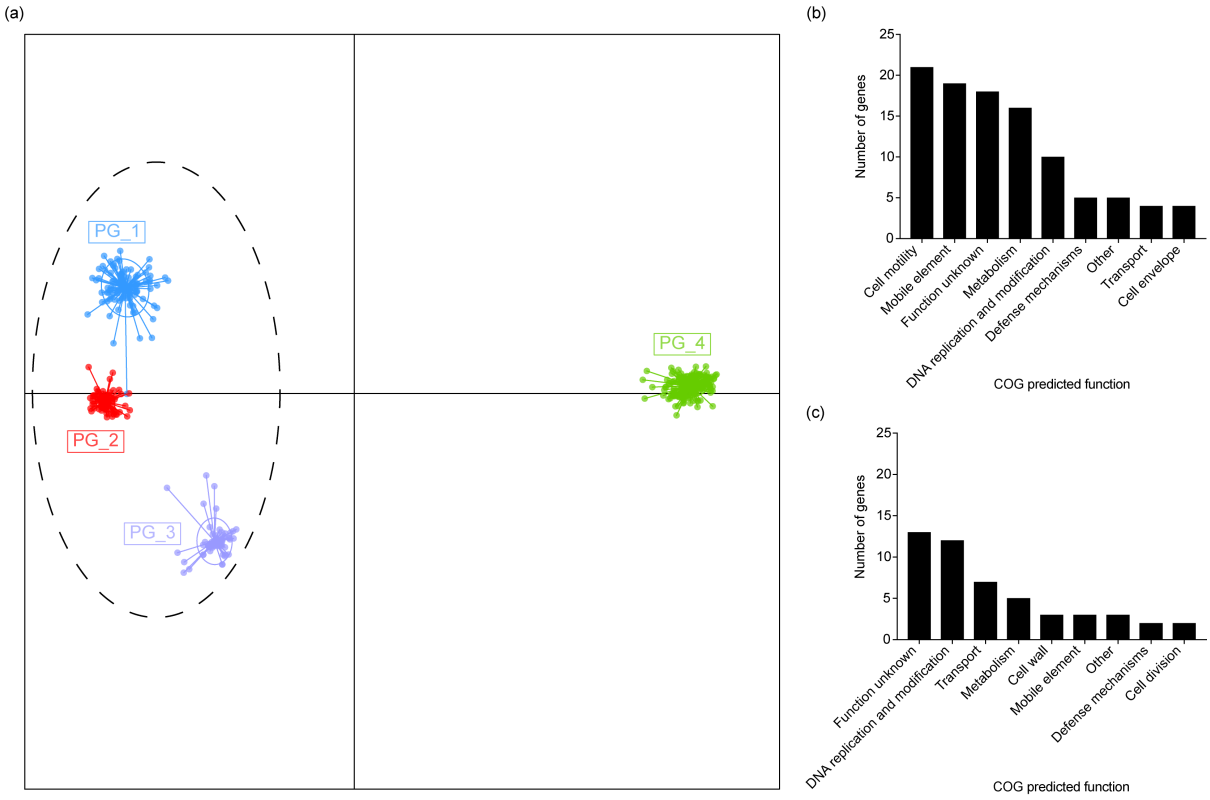
686 *difficile* isolates. Phylogenetic groups of *C. difficile* are shown in circles. Direction of edges

687 represent direction of recombination event (donor to recipient). Range of recombination events

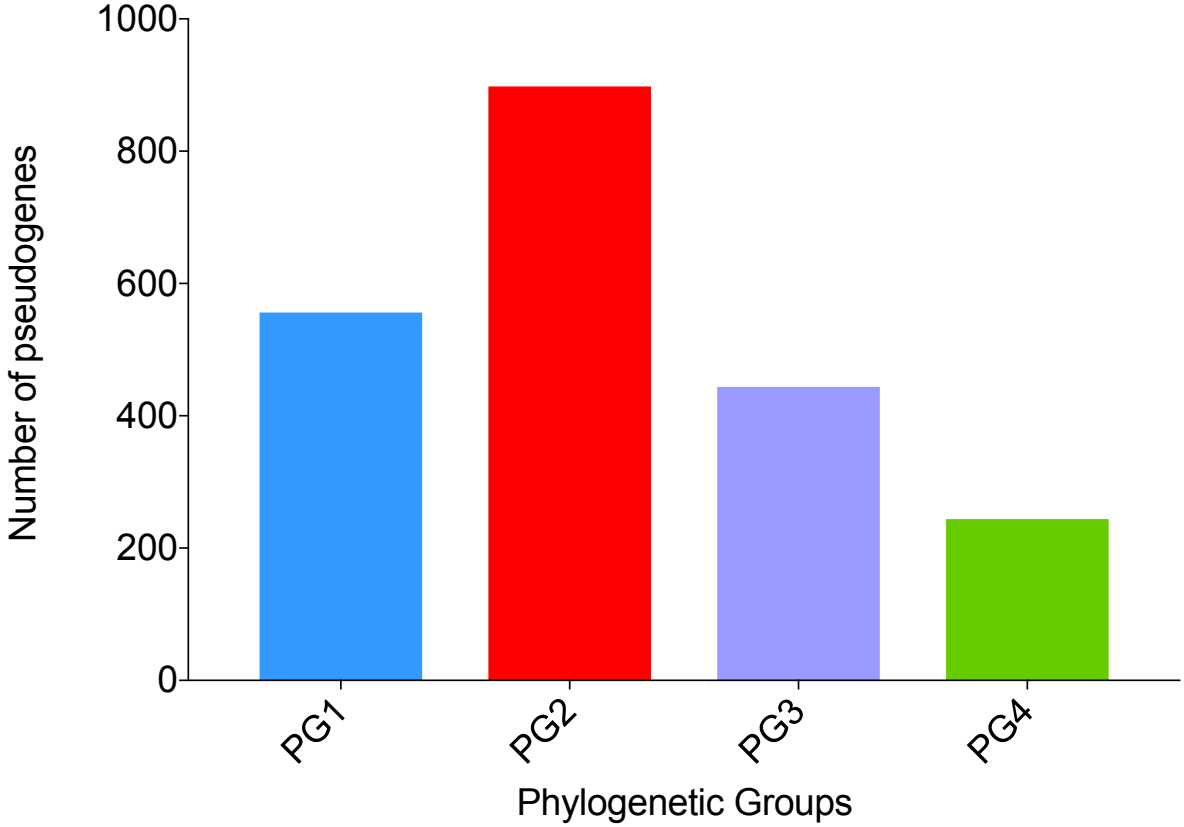
688 are shown on the edges. PG4 represents old *C. difficile* and group of PG1, 2 and 3 represent

689 new *C. difficile.*

690



Supplementary Figure 6. Comparison of accessory genome between 4 phylogenetic groups (PGs) of *Clostridium difficile*. (A) Discriminant analysis of principal components using Clusters of Orthologous Groups (COGs) and accessory genome of strains from PG1 (blue), PG2 (red), PG3 (purple), and PG4 (green). (B) Functional classification and distribution of enriched genes in the group of PG1, 2 and 3 as compared to PG4. (C) Functional classification and distribution of enriched genes in PG4 as compared to the group of PG1, 2 and 3. PG4 represents old *C. difficile* and group of PG1, 2 and 3 represent new *C. difficile*.



706

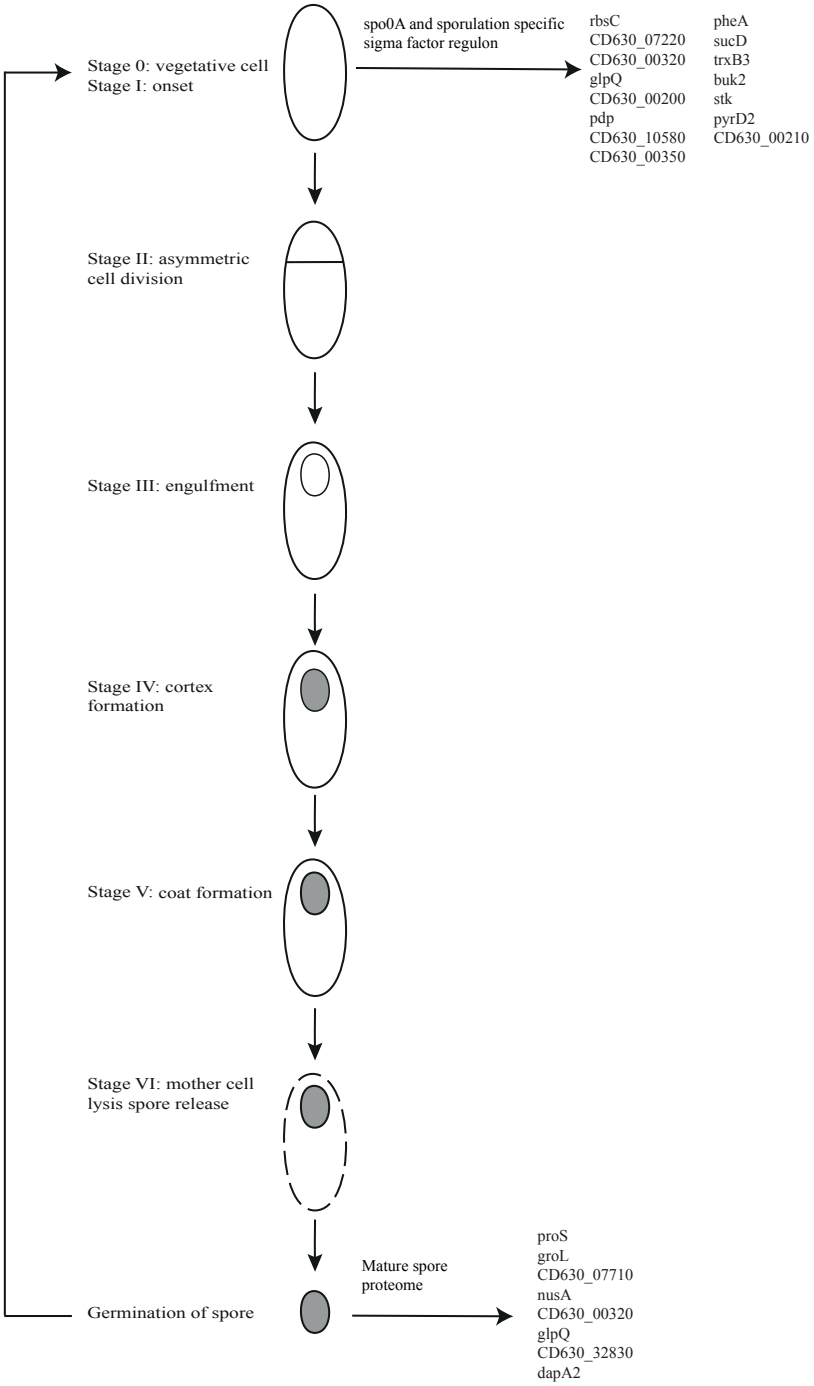
707 **Supplementary Figure 7. High number of pseudogenes in new *Clostridium difficile***

708 **lineage.** The bar-plot shows the number of pseudogenes in each phylogenetic groups (PGs):

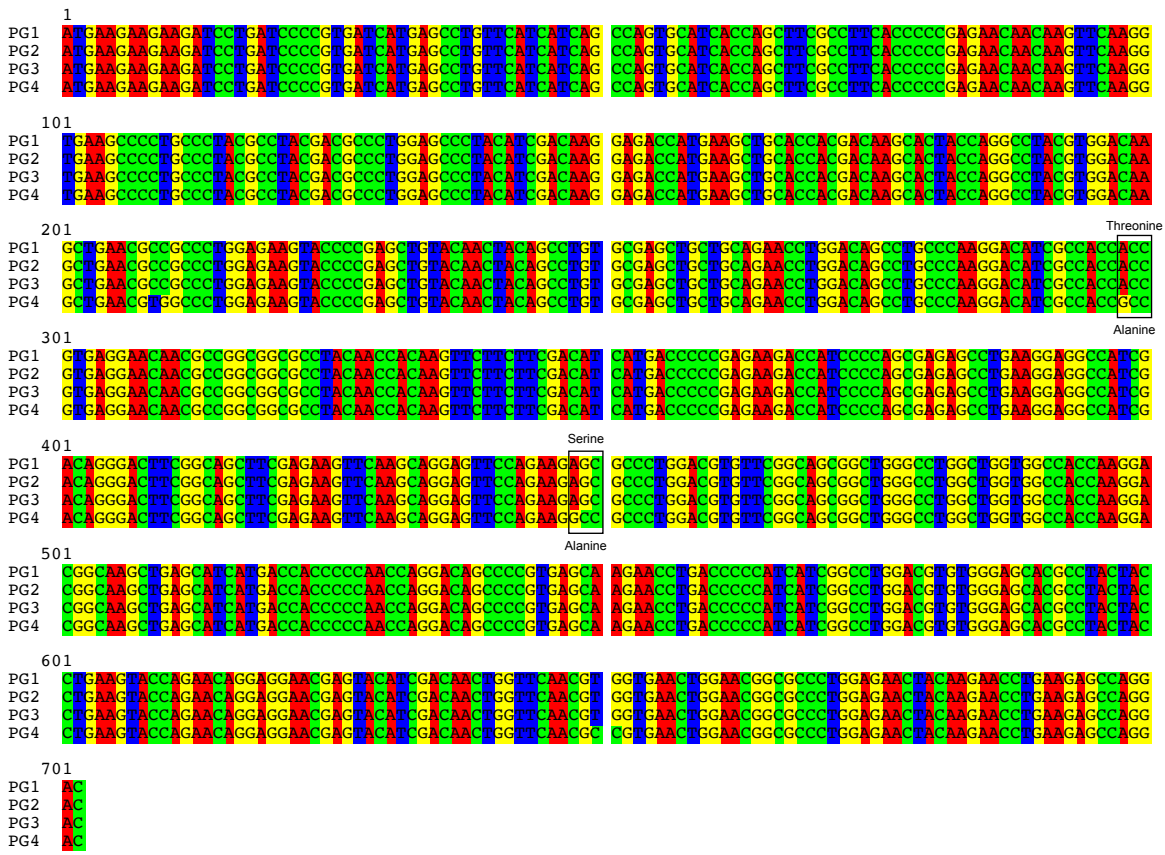
709 PG1 (blue), PG2 (red), PG3 (purple), and PG4 (green). of *Clostridium difficile*. PG4 represents

710 old *C. difficile* and group of PG1, 2 and 3 represent new *C. difficile*.

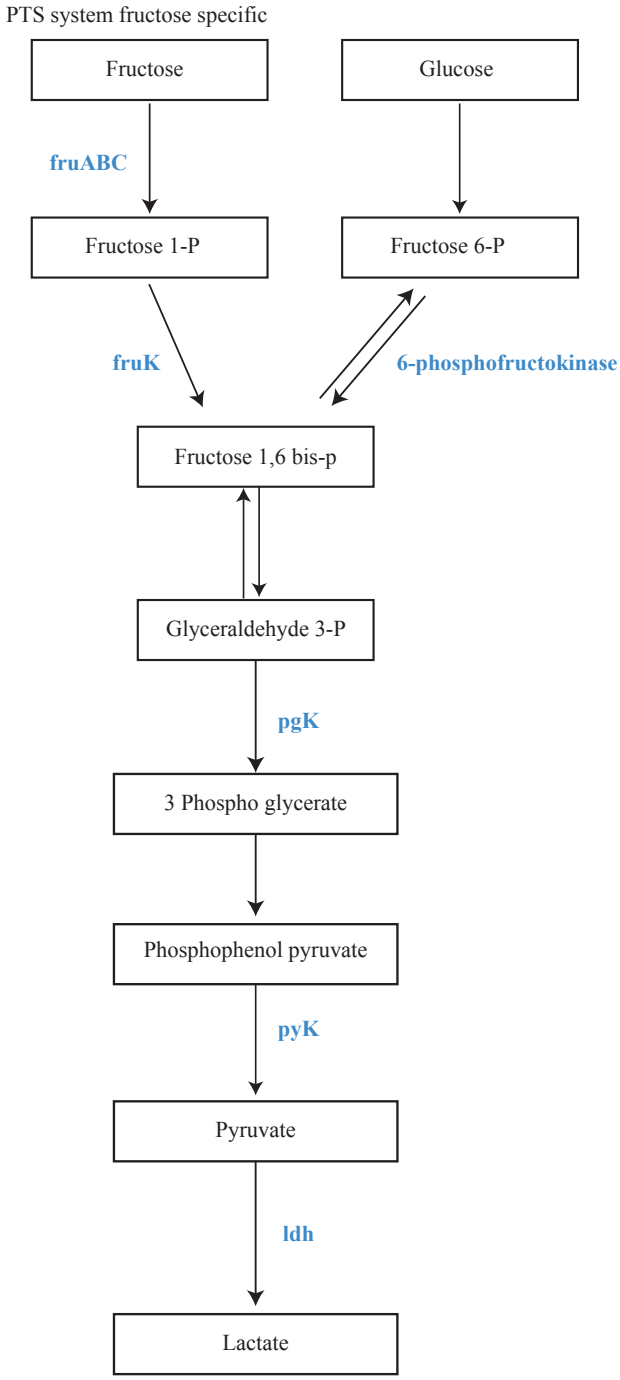
711



Supplementary Figure 8. Sporulation-associated genes in old *Clostridium difficile* lineage. There are 21 sporulation-associated positively selected genes in PG4. These are all either present in the mature spore proteome or they are regulated by Spo0A or its sporulation specific sigma factors. There are no genes directly involved in producing a spore in any of the sporulation stages. PG4 represents old *C. difficile*.



Supplementary Figure 9. Multiple sequence alignment of the *sodA* gene from new and old *Clostridium difficile*. A nucleotide consensus sequence for 4 phylogenetic groups (PG1-4) is shown. Three-point mutations which are present in all new *C. difficile* genomes and absent in old *C. difficile* genomes are shown in black boxes. The amino-acids related to these mutations are mentioned. PG4 represents old *C. difficile* and group of PG1, 2 and 3 represent new *C. difficile*.



725

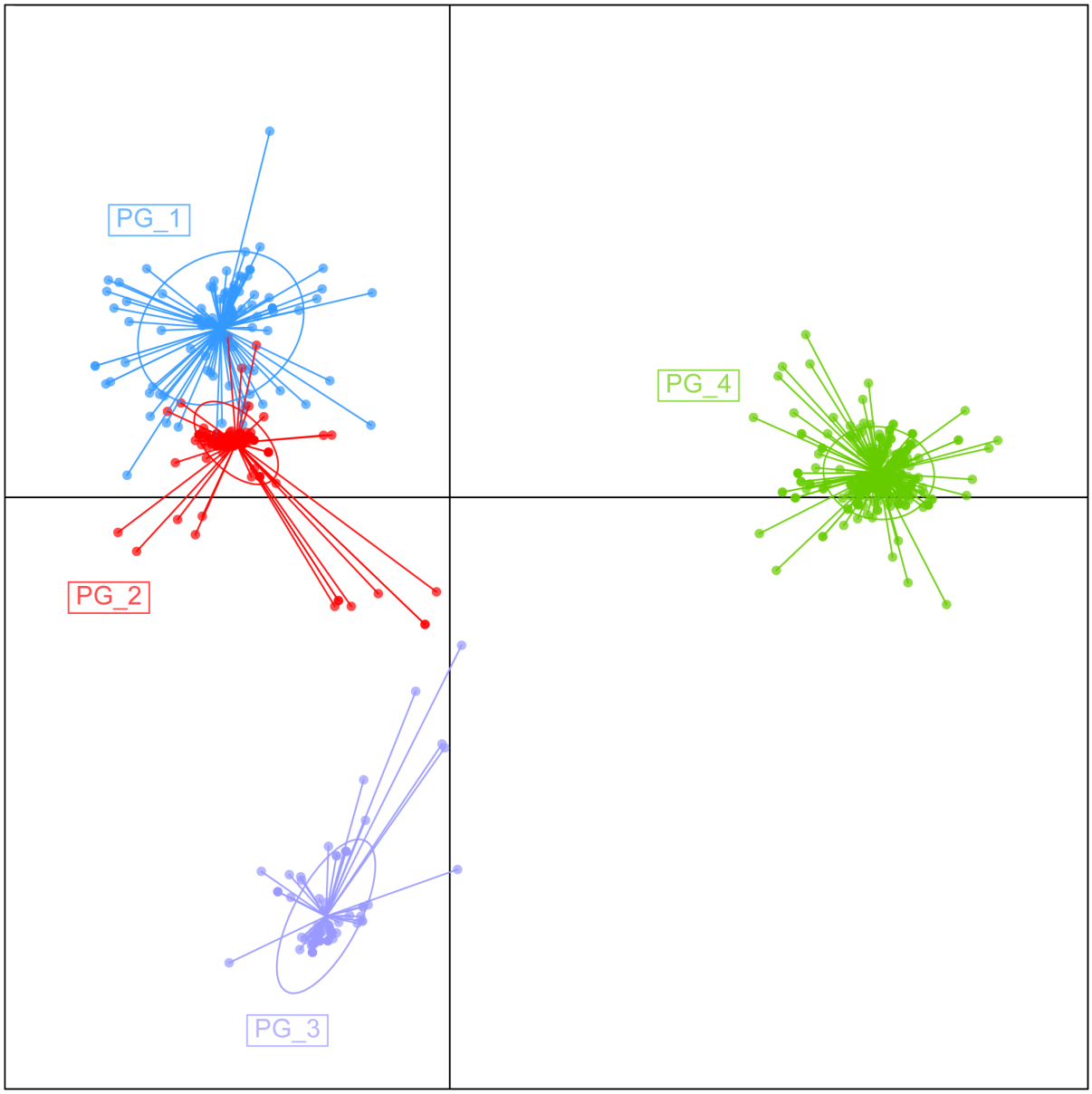
726 **Supplementary Figure 10.** Positively selected genes of new *C. difficile* are shown in blue in

727 the pathway of glucose and fructose in *C. difficile*.

728

729

730



Supplementary Figure 11. Functional diversity of carbohydrate-active enzyme in 4 phylogenetic groups (PGs) of *Clostridium difficile*. Discriminant analysis of principal components using carbohydrate active enzymes (CAZymes) database**.** Each colour represents a strain from 4 PGs: blue (PG1); red (PG2); purple (PG3); and green (PG4). PG4 represents old *C. difficile* and group of PG1, 2 and 3 represent new *C. difficile*.