Title: Strain variation in *Clostridioides difficile* toxin activity associated with genomic variation at both PaLoc and non-PaLoc loci

Running title: *C. difficile* toxin GWAS & evolution

Katie Saund,[a] Ali Pirani,[a,b] D. Borden Lacy[c,d], Philip C. Hanna[a], Evan Snitkin[a,b#]

[a]Department of Microbiology and Immunology, University of Michigan, Ann Arbor, Michigan, USA

[b]Department of Internal Medicine/Division of Infectious Diseases, University of Michigan, Ann Arbor, Michigan, USA

[c]Department of Pathology, Microbiology and Immunology, Vanderbilt University School of Medicine, Nashville, Tennessee, USA

[d]The Veterans Affairs Tennessee Valley Healthcare System, Nashville, Tennessee, USA

#Address correspondence to Evan Snitkin, esnitkin@med.umich.edu

Abstract word count:

Manuscript word count:

ABSTRACT

Clinical disease from *Clostridioides difficile* infection can be mediated by two toxins and their neighboring regulatory genes encoded within the five-gene pathogenicity locus (PaLoc). We provide several lines of evidence that the toxin activity of *C. difficile* may be modulated by genomic variants outside of the PaLoc. We used a phylogenetic tree-based approach to demonstrate discordance between toxin activity and PaLoc evolutionary history, an elastic net method to show the insufficiency of PaLoc variants alone to model toxin activity, and a convergence-based bacterial genome-wide association study (GWAS) to identify correlations between non-PaLoc loci with changes in toxin activity. Combined, these data support a model of

C. difficile disease wherein toxin activity may be strongly affected by many non-PaLoc loci. Additionally, we characterize multiple other *in vitro* phenotypes relevant to human infections including germination and sporulation. These phenotypes vary greatly in their clonality, variability, convergence, and concordance with genomic variation. Lastly, we highlight the intersection of loci identified by GWAS for different phenotypes and clinical severity. This strategy to identify the overlapping loci can facilitate the identification of genetic variation linking phenotypic variation to clinical outcomes.

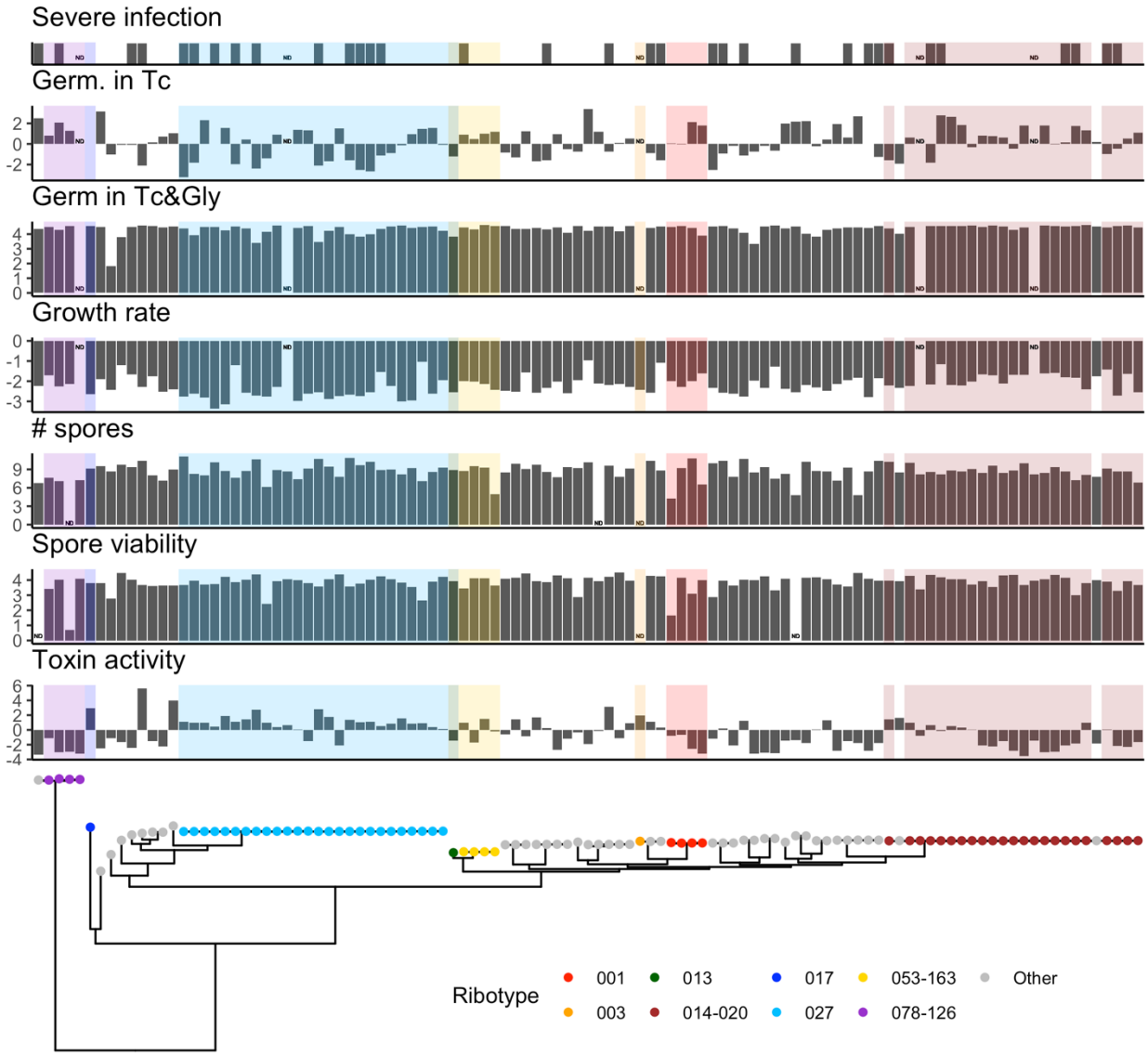
IMPORTANCE

Clostridioides difficile has two major disease mediating toxins, A and B, encoded within the pathogenicity locus (PaLoc). In this study we demonstrate via multiple approaches that genomic variants outside of the PaLoc are associated with changes in toxin activity. These genomic variants may provide new avenues of exploration in the hunt for novel disease modifying interventions. Additionally, we provide insight into the evolution of several additional phenotypes also critical to clinical infection such as sporulation, germination, and growth rate. These *in vitro* phenotypes display a range of responses to evolutionary pressures and as such vary in their appropriateness for certain bacterial genome wide association study approaches. We used a convergence-based association method to identify the genomic variants most correlated with both changes in these phenotypes and disease severity. These overlapping loci may be important to both bacterial function and human clinical disease.

INTRODUCTION

46 *Clostridioides difficile* is a toxin-producing, healthcare-associated bacterial pathogen. It exhibits
47 extensive genetic variation due to its highly mobile genome, a large pangenome, and a most
48 recent common ancestor for clades C1-5 dating back approximately 3.89 million years (1, 2).
49 Such genomic variability has enabled *C. difficile* adaptation to multiple host species and to
50 spread among humans in both nosocomial and community contexts (3). Underlying this genetic
51 variation, is phenotypic variation in many traits including toxin production, sporulation,
52 germination, growth, and virulence (4). This genetic and phenotypic variation has led many to
53 ask whether different genetic backgrounds of *C. difficile* may differ in their propensity to cause
54 severe infections. To this end, many studies have sought to identify key genetic traits harbored
55 by putative hypervirulent strains, such as Ribotype 027 (RT027). Despite this interest and
56 intense study, the genetic basis for variation in phenotypes relevant to the *C. difficile* infection
57 lifecycle remains limited.
58 Disease during *C. difficile* infection is mediated by extracellular toxins, primarily Toxins
59 A (TcdA) and B (TcdB), which damage the cytoskeletons of intestinal cells leading to cell death
60 and gut inflammation. These two toxins are large proteins with four domains:
61 glucosyltransferase, autoprotease, pore-forming, and C-terminal combined repetitive
62 oligopeptides (CROPs) (5). Toxins A and B are both located within the pathogenicity locus
63 (PaLoc) with three other genes: *tcdR*, *tcdC*, and *tcdE*. *tcdR* is a positive regulator of *tcdA* and
64 *tcdB* and encodes an RNA polymerase factor (6). *tcdC* may be a negative regulator of *tcdR* (6).
65 *tcdE* encodes a holin-like protein and may contribute to toxin secretion (7). Many factors and
66 systems are implicated in PaLoc regulation including growth phase, access to specific
67 metabolites, sporulation, quorum sensing, and some flagellar proteins (8). In addition to toxin

68 production, other phenotypes may influence *C. difficile* disease severity or transmission,
69 including sporulation, germination, and growth (9–11).
70 Approaches for uncovering the genomic determinants of bacterial phenotypes, such as
71 toxin activity, include *in vitro* assays, comparative genomics, and bacterial genome-wide
72 association studies (bGWAS). An advantage of bGWAS is the ability to sift through existing
73 genetic variation in bacterial populations to identify variants associated with natural phenotypic
74 variation. In this way, bGWAS can provide insight into phenotypic evolution, and enable the
75 identification of variants of interest that mediate modulation of clinically relevant phenotypes,
76 such as virulence (12). Here, we capitalized on a diverse collection of over 100 *C. difficile*
77 isolates for which multiple phenotypes had previously been characterized (4). We performed
78 whole genome sequencing and used bGWAS to uncover novel genotype-phenotype associations.
79 We explore these genotype-phenotype associations and describe the phenotype variation through
80 phylogenetic and evolutionary analyses. Our analyses reveal the influence of genetic variation on
81 phenotypic variation and help illuminate factors that may be influencing clinical disease.
82



83

84 **FIG 1** Clinical *C. difficile* sample phenotypes aligned with the phylogenetic tree. Color indicates

85 ribotype. ND = No data. *In vitro* phenotypes were log transformed. Infections were classified as

86 severe or not severe.

87

88 **RESULTS**

89 **Distinct evolutionary trajectories of clinically relevant *C. difficile* phenotypes.**

90 To improve our understanding of the evolution of phenotypic diversity in *C. difficile* we
91 performed whole-genome sequencing on a clinical isolate collection that had previously been
92 assayed for toxin activity, two measures of germination, two measures of sporulation, and
93 growth rate (4, 10). Overlaying these phenotypes on a whole-genome phylogeny revealed
94 distinct patterns for each phenotype (Fig. 1). Toxin activity and germination in Tc and Gly are
95 clonal phenotypes that show stable inheritance within lineages, as evidenced by high
96 phylogenetic signal (Fig. 2A). For example, toxin activity displays clonal lineages with
97 uniformly high (e.g. RT027) and low (e.g. RT014) toxin activity (Fig. 1). In contrast,
98 germination in Tc and growth rate are less clonal, with extensive variation even within clonal
99 lineages (Fig. 1,2A). Finally, the two sporulation phenotypes show the least clonality, with
100 virtually no clustering on the phylogeny (Fig. 1,2A). Overall, the range in clonality and
101 phylogenetic signal observed for these phenotypes suggests that despite all being central to the
102 *C. difficile* life cycle, that they are shaped by different evolutionary pressures.
103

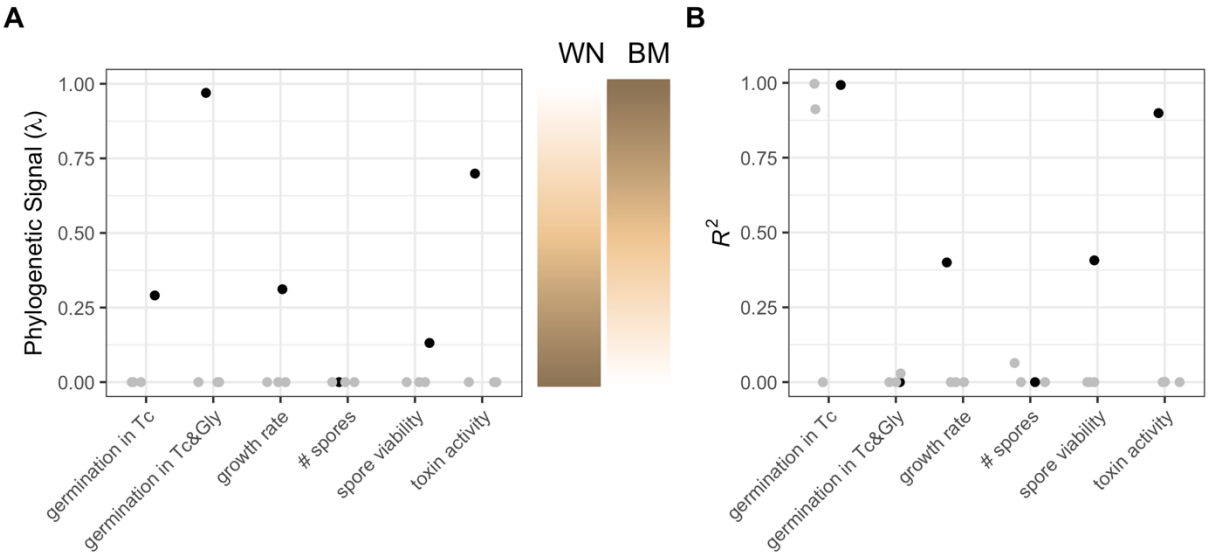


FIG 2 Phenotype phylogenetic signal and genomic model. (A) The phylogenetic signal of each phenotype (black) and its negative controls (grey). WN = white noise. BM = Brownian motion. (B) Elastic nets modeling each phenotype, with high R^2 values indicating that the phenotype is strongly predicted by genetic variation in SNPs. Synonymous SNPs were excluded from this analysis.

In addition to varying in their clonality, the six phenotypes show distinct differences in their overall degree of variation (Table 1). Toxin activity had the largest dispersion with a geometric coefficient of variation of 5.4. The combination of high clonality and high dispersion in toxin activity suggests that *C. difficile* may have evolved multiple successful toxin strategies or have different evolutionary trajectories that are difficult to escape once begun. In contrast, the near uniformity observed in germination in Tc and Gly, could indicate either strong stabilizing selection or inadequate precision of the assay.

119

120

121

122 **TABLE 1** Dispersion (geometric coefficient of variation) and convergence (ratio metric of

123 convergence) of the log transformed phenotypes.

	Germination in Tc	Germination in Tc & Gly	Growth rate	Total spores	Viable spores	Toxin activity
Geometric coefficient of variation	2.8	0.4	0.5	2.2	0.6	5.4
Ratio metric of convergence	46.8	18.0	43.0	38.7	27.3	33.0

124

125 **Phenotypes vary with respect to their association with genetic variation.**

126 Next, we sought to understand the degree to which phenotypic variability in this dataset

127 is genetically encoded. The phenotype best modeled by genomic variants is toxin activity with

128 $R^2 = 0.90$ (Fig. 2B). Growth rate, both sporulation phenotypes, and gemination in Tc and Gly

129 have much lower R^2 values, all $R^2 < 0.50$. Germination in Tc has a high R^2 value, $R^2 = 0.99$, but

130 this finding appears to be spurious as two of the three negative controls using randomly

131 permuted data have similarly high R^2: 0.00, 0.91, and 1.00. The germination and number of

132 spores phenotypes are so poorly encoded by genomic variation that it is suggestive that the

133 assays may lack sufficient precision to capture relevant strain variation, while toxin activity

134 appears far more genetically deterministic.

135

136 **Phenotypes show a range in their level of phylogenetic convergence**

137 A striking feature observed when overlaying the phenotype panel on the whole-genome
138 phylogeny was variation in the frequency of convergence of high or low phenotype values.
139 Convergence, the independent evolution of a trait, may imply the existence of environmental
140 pressures that select towards a specific value or constrain the phenotype's value. To quantify
141 convergence of the different phenotypes we employed the ratio metric, where a higher ratio
142 metric value suggests more episodes of convergence. Germination in Tc has the most
143 convergence, 46.8. The germination in Tc and Gly and spore viability phenotypes have the least
144 convergence, 18.0 and 27.3 respectively. These low values may be driven in part by the lack of
145 dispersion in the phenotype values. The remaining phenotypes demonstrate intermediate levels
146 of phylogenetic convergence. Below we seek to exploit the high level of convergence in certain
147 phenotypes to identify genetic drivers of their variation.
148
149 **Identifying genetic variation associated with phenotypic variation through genome-wide**
150 **association study**
151 Having observed differences in the evolutionary patterns of different phenotypes, we next
152 sought to identify the specific genetic variation that may be underlying phenotypic variation by
153 performing a genome-wide association study (GWAS) for each phenotype. Due to the high
154 convergence in several of the phenotypes (Table 1) and extensive genetic variation in our isolate
155 collection, we opted for a convergence-based GWAS approach that could identify variants of
156 interest by their non-random co-convergence with a phenotype. The genotypes tested included
157 approximately 69,600 SNPs, 8,400 indels, and 7,500 accessory genes. Significantly associated
158 variants were identified for growth rate, number of spores, toxin activity, germination in Tc, and
159 severity.

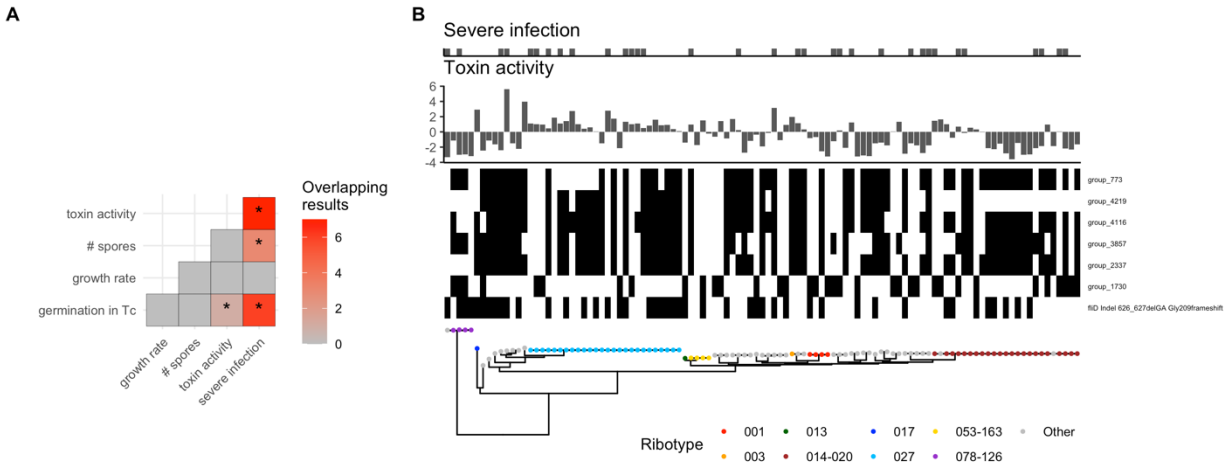
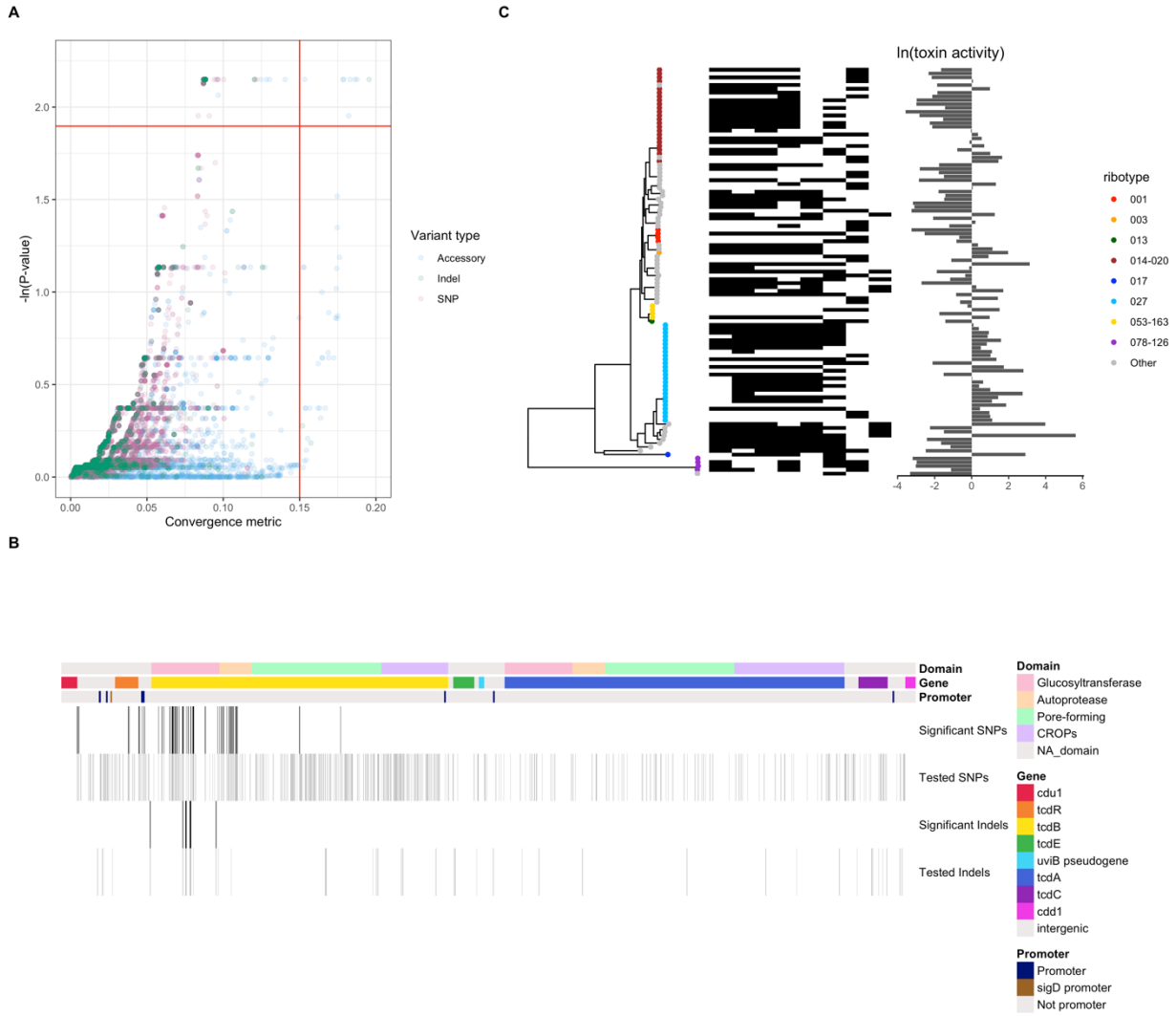


FIG 3 Overlapping GWAS results. (A) Heatmap indicates the number of shared GWAS results with significant *P*-values and high levels of convergence in the Continuous Test (continuous phenotypes) or Synchronous Test (Severity). Asterisks indicate significantly more overlapping results than expected by chance ($P < 0.05$). The two phenotypes lacking any GWAS results with significant *P*-values and high levels of convergence were excluded. (B) Shared hits between the toxin activity and severe infection GWAS. Top: phenotype. Center: heatmap indicating the presence of loci with both significant *P*-values and high levels of convergence in both toxin activity and severity GWAS results. Bottom: phylogenetic tree labeled by ribotype.

Overlapping GWAS results. Despite the phenotypes showing distinct evolutionary patterns, we first explored whether there was evidence of overlap in the genetic circuits modulating the different traits. We cataloged the extent of this overlap by counting the number of intersecting genomic loci with both high significance and convergence in each pair of GWAS results. Three of the four phenotypes shared more hits with the severe infection GWAS results

175 than expected by chance via a permutation test (Fig. 3A). Toxin activity and severe infection
176 have the most overlap with 7 shared loci. These shared loci include six accessory genes and a
177 frameshift mutation at Glycine 209 in flagellar hook-associated protein 2 (*fliD)* (Fig. 3B). The
178 *fliD* finding is consistent with known co-regulation that occurs between flagellar and toxin
179 systems in *C. difficile* that is mediated in part by SigD, a sigma factor that binds to a *tcdR*
180 promoter region and positively regulates *tcdR* (13).



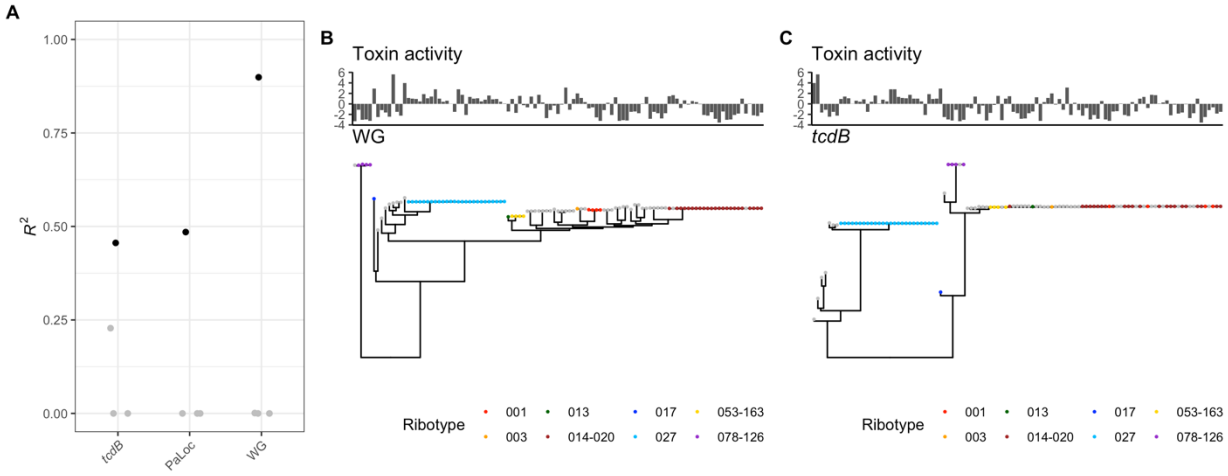
181

182 **FIG 4** Genome-wide association study identified genomic variants associated with toxin activity
183 variation. (A) GWAS results. Tested loci are either accessory genes (blue; N=4,352), SNPs

184 (pink; N=12,167), or indels (green; N=1,843). The red horizontal line indicates a False
185 Discovery Rate of 15%. The red vertical line separates low vs high convergence. (B)
186 Significantly associated loci from GWAS located in the PaLoc. Of the 633 PaLoc variants (SNP
187 N=563, Indel N=70) tested by GWAS only the variants significantly associated are plotted as
188 vertical bars (SNP N=71, Indel N=16). Top annotation: toxin protein domains in *tcdB*. Center
189 annotation: gene. Bottom annotation: promoter locations. (C) Left: phylogenetic tree labeled by
190 ribotype. Center: heatmap indicating the presence of loci significantly associated with toxin
191 activity and with high convergence. Right: toxin activity.
192
193 **Genetic variation associated with modulation of toxin activity**
194 For the remainder of our analysis, we focused on understanding genetic variation associated with
195 variation in toxin activity. In addition to the central role of toxin in *C. difficile* disease, our
196 decision to focus on toxin was motivated by it being the phenotype being best explained by
197 genetic variation in sequenced strains (Figure 2B). In the following sections we examine variants
198 playing a key role in modulating toxin activity.
199 The toxin activity GWAS identified many genomic variants both significantly associated
200 with toxin activity changes and had high levels of convergence (Fig. 4A). As the PaLoc encodes
201 toxin genes and regulators we expected that variants located within the PaLoc would be
202 significantly associated with toxin activity and used this as a positive control for our analysis.
203 Consistent with this, we observed PaLoc variants in the pool of significant results associated
204 with toxin activity. Eighty-seven of the 220 loci significantly associated with toxin activity occur
205 in the PaLoc. Given that the toxin activity assay used is a measure of Toxin B activity it is
206 particularly reassuring that these 87 PaLoc loci include 75 *tcdB* variants and 2 *tcdR-tcdB*

207 intergenic region variants (Fig. 4B). Indeed, these variants are a significant enrichment compared
208 to the number of variants within or flanking *tcdB* that are expected by chance using a
209 permutation approach, $P = 0.0001$ (median = 1; range = 0-10). *tcdB* variants were found in all
210 four protein domains, but the significantly associated variants are mostly found within the
211 glucosyltransferase and autoprotease domains (Fig. 4B). Certain significant missense variants
212 within *tcdB* have plausible functional impacts on Toxin B such as an adenosine to cytosine
213 transversion at position 1967 which changes an aspartic acid to alanine; this mutation occurs near
214 the zinc binding site and could theoretically affect toxin autoprocessing within the host cell. Of
215 the 15 tested variants that occur within the *tcdR-tcdB* intergenic region, 5 were significantly
216 associated with toxin activity. Three of these variants occur within a *tcdB* promoter suggesting a
217 potential role in modulating sigma factor binding and therefore altering *tcdB* transcription. A
218 notable lack of association was observed between an adenosine deletion at nucleotide 117 in
219 *tcdC* that has been suggested to cause increased toxin production in RT027 (14). This deletion
220 was found in all 26 RT027 samples as well as in 3 additional samples ("Other" ribotype) but did
221 not reach significance in the GWAS, P=0.95.
222 Next, we sought to generate hypotheses about new associations between genomic
223 variants and toxin activity that reside outside of the PaLoc. The variants that were significant and
224 had high ε, a metric of shared genotype-phenotype convergence, are cataloged in File S1 and
225 plotted in Fig. 4C (15). A single ε value captures the number of tree edges where both a genotype
226 is mutated and the toxin activity value has a large change. ε values close to zero suggest that the
227 genotype mutates on very few edges where the toxin activity changes drastically. The loci
228 associated with changes in toxin activity are present in multiple, independent lineages (Fig. 4C).
229 The previously mentioned frameshift mutation in *fliD* is the variant most strongly associated

230 with changes in toxin activity when ranked by ε then *P*-value. The next most strongly associated
231 variant maps to CD630_02364 which is annotated as a putative signaling protein in the reference
232 genome CD630. The other most strongly associated variants are poorly annotated accessory
233 genes. The significant accessory genes identified by this analysis may yield profitable results in
234 mechanistic studies dissecting *C. difficile* toxin activity and could be prioritized for further
235 characterization.
236
237
238



239
240 **FIG 5** *tcdB* variation does not fully model toxin activity. (A) Elastic net model performance of
241 toxin activity. Models were built from *tcdB* variants, PaLoc variants, or whole genome (WG)
242 variants. Toxin activity with a tree built from (B) the whole genome or (C) just *tcdB*.
243
244 **Genetic variation at PaLoc accounts for only half of phenotypic variation in toxin activity.**
245 The GWAS identified both PaLoc and non-PaLoc loci correlated with variation in toxin
246 activity. To understand the relative contribution of genetic variation in the PaLoc to variation in

247 toxin activity we employed an elastic net approach. Models of toxin activity constructed with
248 different subsets of variants found that PaLoc variants and *tcdB* variants have similar abilities to
249 model toxin activity, $R^2 = 0.48$ and $R^2 = 0.46$ respectively (Fig. 5A). However, variants from the
250 whole genome build a more accurate model of toxin activity, $R^2 = 0.90$ (Fig. 5A). Of the 634
251 variants in the PaLoc, 404 (64%) occur in *tcdB* or in its flanking intergenic regions; in the best
252 performing elastic net model derived from PaLoc variants, 34/61 (56%) of the variants are
253 mutations in *tcdB* or its flanking regions. In the whole genome model only 17/1795 (1%) of
254 variants occur in *tcdB* or its flanking regions.

255 To assess the predictive capacity of PaLoc variation in a different way, we compared
256 phylogenetic trees built from whole genome variants and the *tcdB* gene. As there is far less
257 variation in *tcdB* than the whole genome, we observe many polytomies in the *tcdB* gene tree and
258 none in the whole genome tree (Fig 5B,C). While the *tcdB* gene tree's toxin activity is better
259 modeled by Brownian motion, $\lambda = 0.94$, than in the whole genome tree, $\lambda = 0.75$ (Fig. 2A), there
260 remains much toxin activity variation unexplained by tree structure. Given the unexplained toxin
261 activity variation on the *tcdB* gene tree and variation not captured in the toxin activity elastic net
262 model, we conclude that while *tcdB* gene variation is likely an important mediator in the
263 evolution of toxin activity, other loci play a key role as well. Finally, the whole genome model
264 suggests that many loci besides *tcdB* may affect *C. difficile* toxicity and therefore a wider lens for
265 examining genetic influences on toxicity will be fruitful.

266

267 **DISCUSSION**

268 *C. difficile* is a genetically diverse pathogen, with extensive variation in both its core and
269 accessory genome. Currently, we have a limited understanding of the functional impact of most

270 of this variation and how it relates to *C. difficile* infection. Here, we attempted to improve our
271 understanding of the genotype to phenotype map in *C. difficile* by analyzing variation in
272 clinically relevant phenotypes in the context of *C. difficile* genomic variants. We observe that
273 despite their central role to the *C. difficile* transmission and infection cycle sporulation,
274 germination, growth and toxin activity show distinct evolutionary trajectories. Focusing on the
275 phenotype thought to be most closely linked to virulence, we observe that toxin activity is highly
276 clonal, with lineages tending to either possess high or low toxin activity. Consistent with prior
277 reports we find that variation in toxin activity can be modulated by variants in the PaLoc,
278 however we find that more than 50% of phenotypic variation is associated with genetic variation
279 outside of the PaLoc.

280 Our exploration of these *C. difficile* phenotypes revealed a broad range of clonality,
281 dispersion, association with genomic variation, and convergence. As such, each phenotype
282 appears to be shaped by different selection forces. The existence of phenotypes that show no
283 association with the recombination filtered phylogeny could indicate either a lack of precision in
284 the laboratory assay or a strong role for recombinant genomic regions in shaping these
285 phenotypes. We focused our analysis on toxin activity, in part, because of the precision of the *in*
286 *vitro* assay results and its high degree of genetic determinism. Regardless of the basis for the lack
287 of phylogenetic signal in some of the non-toxin phenotypes, these results show how overlaying
288 phenotypic variation on whole-genome phylogenies provides useful context for interpreting and
289 scrutinizing experimental measurements, and in this case clearly demonstrates the rich and
290 varied patterns of evolution among *C. difficile* strains.

291 Toxigenic bacterial species that require live transmission may undergo strong selective
292 pressure to promote host survival and therefore bias towards lower toxin activity (16). In

293 contrast, sporogenic *C. difficile* can survive and transmit even after the host dies; this may reduce
294 the strength of selection on toxin activity and therefore many different toxin strategies are
295 successful. Indeed, there are prolific toxigenic and non-toxigenic strains of *C. difficile*.
296 Additionally, the species has had multiple independent losses of the PaLoc (17), with our results
297 indicating that even strains harboring an intact PaLoc may evolve to have decreased toxin
298 activity. The *C. difficile* strains with high toxin activity may have success by shaping a hostile
299 metabolic state in the host gut that these bacteria are able to uniquely exploit (18) or its more
300 severe, inflammatory infection which results in diarrhea and therefore increased transmission.
301 This then raises the question of what the selective pressure for lower toxin activity may be. One
302 possibility is that toxin activity itself may not be the most critical aspect of the toxin upon which
303 evolution is acting, with other aspects such as toxin immunogenicity potentially evoking a
304 stronger selection pressure. Toxin that evades immune recognition could lead to longer
305 infections and therefore increased transmission, so the strongest selective pressure may be at the
306 surface domains of the toxin proteins rather than on regulators of toxin activity (19). For
307 example, we observed multiple missense variants on the surface of tcdB in this isolate collection,
308 including a glutamic acid 329 to glycine missense variant and threonine 430 to alanine.
309 Our study has several important limitations. First, the limited sample size of this *C.*
310 *difficile* collection could lead to underreporting of clonality of some phenotypes for
311 underrepresented ribotypes and limits power to detect variation with smaller phenotypic impacts.
312 Second, many genomic features such as copy number variants, large structural variants, and
313 plasmids were not included in our GWAS or elastic net models, therefore these analyses are
314 missing some genome encoded information. Similarly, we did not consider the complexity of
315 epistatic interactions between genomic variants on phenotypes.

316 A replication study in a second *C. difficile* cohort in which the toxin activity assay and
317 GWAS is repeated could help prioritize the genomic variants more likely to be causal of changes
318 in toxin activity. The loci identified in both this study and the proposed study would be higher
319 confidence candidates for experiments that examine the effect of those potential variants on toxin
320 activity. Additional studies investigating *C. difficile in vitro* phenotypes from an evolutionary
321 perspective would help to prioritize the phenotypes that may offer the most insight into the
322 success and regulation of certain strains.
323

324 **MATERIAL AND METHODS**

325 **Study population and *in vitro* characterization.** The University of Michigan Institutional
326 Review Board approved all sample and clinical data collection protocols used in this study
327 (HUM00034766). Where applicable, written, informed consent was received from all patients
328 prior to inclusion in this study. Stool samples were collected from a cohort of 106 Michigan
329 Medicine patients with *C. difficile* infection from 2010-2011, which included all severe cases
330 during the study period (4, 10). Cases were classified as severe if the infection required ICU
331 admission or interventional surgery, or if the patient died within 30 days of infection diagnosis.
332 A clonal spore stock from each patient was used for ribotyping and *in vitro* studies. Previous
333 experiments characterized the germination in taurocholate (TC; %), and germination in Tc and
334 glycine (Gly; %), maximum growth rates (OD_{600}/hour), total spore production (heat resistant
335 colon forming units per ml), viable spores (%), and equivalent toxin B activity (ng/ml) (4, 10).
336 Taurocholate is a physiologic bile salt known to cause *C. difficile* germination; glycine can
337 increase germination with taurocholate (20). Samples were classified as severe infections if they

338 were collected from a patient whose *C. difficile* infection required ICU admission or
339 interventional surgery, or if the patient died within 30 days of infection diagnosis (4, 10).
340 **Genomic analysis**. The spore stocks were grown in an anaerobic chamber overnight on
341 taurocholate-coition-cycloserine-fructose agar plates. The next day a single colony of each
342 sample was picked and grown in Brain Heart Infusion medium with yeast extract liquid culture
343 media overnight. The vegetative *C. difficile* cells were pelleted by centrifugation, washed, and
344 then total genomic DNA was extracted. Genomic DNA extracted with the MoBio PowerMag
345 Microbial DNA Isolation Kit (Qiagen) from *C. difficile* isolates (N=108) was prepared for
346 sequencing using the Illumina Nextera DNA Flex Library Preparation Kit. Sequencing was
347 performed on either an Illumina HiSeq 4000 System at the University of Michigan Advanced
348 Genomics Core or an Illumina MiSeq System at the University of Michigan Microbial Systems
349 Molecular Biology Laboratories. Quality of reads was assessed with FastQC v0.11.9 (21).
350 Adapter sequences and low-quality bases were removed with Trimmomatic v0.36 (22). Variants
351 were identified by mapping filtered reads to the CD630 reference genome (GenBank accession
352 number AM180355.1) using bwa v0.7.17 (23), removing polymerase chain reaction duplicates
353 with Picard 2.21.7 (24), removing clipped alignments using Samclip 0.4.0 (25), and calling
354 variants with SAMtools v1.11and bcftools (26). Variants were filtered from raw results using
355 GATK's VariantFiltration v3.8 (QUAL, >100; MQ, >50; >=10 reads supporting variant; and FQ,
356 <0.025) (27). SNPs and indels were referenced to the ancestral allele using snitkitr v0.0.0.9000
357 (28). Pangenome analysis was performed with roary (29). Accessory genes annotations were
358 assigned by prokka v1.14.5 (30).

359 **Data availability**. Sequence data are available under Bioproject PRJNA594943. Details
360 on sequenced strains are available in File S2. Sequences for genes identified by roary are
361 available in File S3.

362 **Phylogenetic analysis.** Consensus files generated during variant calling were
363 recombination filtered using Gubbins v3.0.0 (31). The alleles at each position that passed
364 filtering were concatenated to generate a non-core variant alignment relative to the CD630
365 reference genome. Alleles that did not pass filtering were considered unknown (denoted as N in
366 the alignment). Variant positions in the alignment were used to reconstruct a maximum
367 likelihood phylogeny with IQ-TREE v1.5.5 using ultrafast bootstrap with 1,000 replicates (32,
368 33). ModelFinder limited to ascertainment bias-corrected models was used to identify the best
369 nucleotide substitution model (34). The tree was midpoint rooted. The *tcdB* multiple sequence
370 alignment was built by PRANK v.170427 using only the *tcdB* gene and the resulting tree was
371 midpoint rooted (35). The trees are available in Files S4 and S5.

372 **Genome-wide association studies.** GWAS were performed with hogwash v1.2.4 (15).
373 Phenotype data were natural log transformed. Hogwash settings: bootstrap threshold=0.95,
374 permutations=10,000, false discovery rate=15%. The analysis included SNPs, indels, and
375 accessory genes. The intersection of hogwash results was restricted to results with $\varepsilon > 0.15$ and
376 P-value < 0.15. Only SNPs classified as having "Moderate", "High", or "Modifier" impact by
377 SnpEff v4.3.1 were included (36).

378 **Data analysis.** Data analysis with R v3.6.2 (37) was performed with following packages:
379 ape v5.3 (38), aplot v0.0.6 (39), data.table v1.12.8 (40), ggtree v2.0.4 (41), ggpubr v0.4.0 (42),
380 pheatmap v1.0.12 (43), phytools v0.6-99 (44), tidyverse v1.3.0 (45). Conda v4.9.2 was used to

maintain working environments (46). Analysis code is available at https://github.com/katiesaund/cdifficile_gwas.

Permutation testing. The empirical *P*-value for enrichment of toxin variants in the significant GWAS results and shared results in the overlapping GWAS section were generated via permutation testing. This approach generates a *P*-value by comparing the observed number of events in the data to a distribution of the number of events simulated under the null hypothesis. The null distribution was generated from random sampling without replacement repeated in 10,000 trials (toxin variants) or 1,000 trials (overlapping hits). Multiple testing correction was applied to the overlapping hits analysis using Bonferroni correction.

Convergence analysis. We calculated the degree of convergence of each phenotype using the ratio method (47), which is the ratio of two sample's pairwise patristic distance divided by their pairwise phenotypic distance. We report the average of the scaled pairwise branch length distance (patristic distance) divided by scaled pairwise phenotypic distance for each phenotype. A high value suggests an episode of convergence.

Geometric coefficient of variance. We calculated the dispersion of each phenotype as defined by the geometric coefficient of variance: $\sqrt{\boldsymbol{e}^{\sigma^2} - 1}$ where σ is the standard deviation of the log transformed data.

Phylogenetic signal. Phylogenetic signal is a metric that captures the tendency for closely related samples on a tree to be more similar to each other than they are to random samples on the tree. We calculated phylogenetic signal for each continuous phenotype using Pagel's λ (48). Note that a phenotype that is modeled well by Brownian motion has a λ near 1 while a white noise phenotype has a λ near 0 (48). Negative controls for the phenotypes were created by randomly redistributing each phenotype on the tree.

404 **Elastic net modeling.** We calculated the degree of genetic encoding of each phenotype
405 by modeling a phenotype from genomic variants using elastic net regularization as implemented
406 by pyseer. Pyseer v1.3 command line arguments: --wg enet --n-fold 10 (49). SNPs, indels and
407 accessory genes were all used to model all continuous phenotypes. For all elastic net models only
408 SNPs classified as having "Moderate", "High", or "Modifier" impact by SnpEff were included
409 (36). Toxin activity was additionally modeled by 1) a model built from just PaLoc SNPs and
410 indels and 2) a model built from just *tcdB* SNPs and indels. To determine the value of α, a
411 parameter which controls the ratio of L1 and L2 regularization in the model, five α values were
412 tested for each model: 0.01, 0.245, 0.500, 0.745, and 0.990. The model results with the highest
413 R^2 value were reported. The best α for models of germination in Tc, germination in Tc and Gly,
414 total spores, and toxin activity (all variants) is 0.01. The best α for models of viable spores and
415 growth rate is 0.245. The best α to model toxin activity (*tcdB*) is 0.500. The best α to model
416 toxin activity (PaLoc) is 0.745. Negative controls for the phenotypes were created by randomly
417 redistributing each phenotype on the tree.
418
419 **SUPPLEMENTAL MATERIAL**
420 File S1: Toxin GWAS results. A table with variant name, p-value, and epsilon value.
421 File S2: Bioproject details for the sequenced strains.
422 File S3: Sequences of the genes identified by roary.
423 File S4: Phylogenetic tree.
424 File S5: *tcdB* gene phylogenetic tree.
425
426

427 **ACKNOWLEDGEMENTS**

428 K. S. was supported by the National Institutes of Health (T32GM007544). E. S. S. and A. P.
429 were supported by the National Institutes of Health (1U01AI124255). Work in the Lacy lab is
430 supported by NIH AI095755 and VA BX002943.

431 **REFERENCES**

432 1. Mullany P, Allan E, Roberts AP. 2015. Mobile genetic elements in Clostridium difficile
433 and their role in genome function. Res Microbiol 166:361–367.
434 2. Knight DR, Imwattana K, Kullin B, Guerrero-Araya E, Paredes-Sabja D, Didelot X,
435 Dingle KE, Eyre DW, Rodríguez C, Riley T V. 2021. Major genetic discontinuity and
436 novel toxigenic species in Clostridioides difficile taxonomy. Elife 10:1–25.
437 3. Knight DR, Elliott B, Chang BJ, Perkins TT, Riley T V. 2015. Diversity and evolution in
438 the genome of Clostridium difficile. Clin Microbiol Rev 28:721–741.
439 4. Carlson PE, Walk ST, Bourgis AET, Liu MW, Kopliku F, Lo E, Young VB, Aronoff DM,
440 Hanna PC. 2013. The relationship between phenotype, ribotype, and clinical disease in
441 human Clostridium difficile isolates. Anaerobe 24:109–116.
442 5. Pruitt RN, Lacy DB. 2012. Toward a structural understanding of Clostridium difficile
443 toxins A and B. Front Cell Infect Microbiol 2:28.
444 6. Monot M, Eckert C, Lemire A, Hamiot A, Dubois T, Tessier C, Dumoulard B, Hamel B,
445 Petit A, Lalande V, Ma L, Bouchier C, Barbut F, Dupuy B. 2015. Clostridium difficile:
446 New Insights into the Evolution of the Pathogenicity Locus. Sci Rep 5.
447 7. Govind R, Dupuy B. 2012. Secretion of Clostridium difficile Toxins A and B Requires the
448 Holin-like Protein TcdE. PLoS Pathog 8:e1002727.
449 8. Martin-Verstraete I, Peltier J, Dupuy B. 2016. The regulatory networks that control

450 Clostridium difficile toxin synthesis. Toxins (Basel). Multidisciplinary Digital Publishing
451 Institute.
452 9. Burns DA, Minton NP. 2011. Sporulation studies in Clostridium difficile. J Microbiol
453 Methods 87:133–138.
454 10. Carlson PE, Kaiser AM, Mccolm SA, Bauer JM, Young VB, Aronoff DM, Hanna PC.
455 2015. Variation in germination of Clostridium difficile clinical isolates correlates to
456 disease severity. Anaerobe 33:64–70.
457 11. Tschudin-Sutter S, Braissant O, Erb S, Stranden A, Bonkat G, Frei R, Widmer AF. 2016.
458 Growth Patterns of Clostridium difficile - Correlations with Strains, Binary Toxin and
459 Disease Severity: A Prospective Cohort Study. PLoS One 11:e0161711.
460 12. Laabei M, Recker M, Rudkin JK, Aldeljawi M, Gulay Z, Sloan TJ, Williams P, Endres JL,
461 Bayles KW, Fey PD, Yajjala VK, Widhelm T, Hawkins E, Lewis K, Parfett S, Scowen L,
462 Peacock SJ, Holden M, Wilson D, Read TD, Van Den Elsen J, Priest NK, Feil EJ, Hurst
463 LD, Josefsson E, Massey RC. 2014. Predicting the virulence of MRSA from its genome
464 sequence. Genome Res 24:839–849.
465 13. El Meouche I, Peltier J, Monot M, Soutourina O, Pestel-Caron M, Dupuy B, Pons JL.
466 2013. Characterization of the SigD regulon of C. difficile and its positive control of toxin
467 production through the regulation of tcdR. PLoS One 8:1–17.
468 14. Warny M, Pepin J, Fang A, Killgore G, Thompson A, Brazier J, Frost E, McDonald LC.
469 2005. Toxin production by an emerging strain of Clostridium difficile associated with
470 outbreaks of severe disease in North America and Europe. Lancet 366:1079–1084.
471 15. Saund K, Snitkin ES. 2020. Hogwash: three methods for genome-wide association studies
472 in bacteria. Microb Genomics https://doi.org/10.1099/mgen.0.000469.

473 16. Rudkin JK, McLoughlin RM, Preston A, Massey RC. 2017. Bacterial toxins: Offensive,
474 defensive, or something else altogether? PLoS Pathog 13:1–12.
475 17. Mansfield MJ, Tremblay BJM, Zeng J, Wei X, Hodgins H, Worley J, Bry L, Dong M,
476 Doxey AC. 2020. Phylogenomics of 8,839 Clostridioides difficile genomes reveals
477 recombination-driven evolution and diversification of toxin A and B. PLoS Pathog 16:1–
478 24.
479 18. Fletcher JR, Pike CM, Parsons RJ, Rivera AJ, Foley MH, McLaren MR, Montgomery SA,
480 Theriot CM. 2021. Clostridioides difficile exploits toxin-mediated inflammation to alter
481 the host nutritional landscape and exclude competitors from the gut microbiota. Nat
482 Commun 12:1–14.
483 19. Mansfield MJ, Tremblay BJ-M, Zeng J, Wei X, Hodgins H, Worley J, Bry L, Dong M,
484 Doxey AC. 2020. Phylogenomics of 8,839 Clostridioides difficile genomes reveals
485 recombination-driven evolution and diversification of toxin A and B. Biorvix.
486 20. Sorg JA, Sonenshein AL. 2008. Bile salts and glycine as cogerminants for Clostridium
487 difficile spores. J Bacteriol 190:2505–2512.
488 21. Andrews Si. 2010. FastQC: a quality control tool for high throughput sequence data.
489 22. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina
490 sequence data. Bioinformatics 30:2114–2120.
491 23. Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler
492 transform. Bioinformatics 25:1754–1760.
493 24. Broad Institute. Picard Tools.
494 25. Seemann T. samclip.
495 26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G,

496 Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics
497 25:2078–2079.
498 27. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K,
499 Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: A
500 MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res
501 20:1297–1303.
502 28. Saund K, Lapp Z, Thiede SN. snitkitr.
503 29. Page AJ, Cummins CA, Hunt M, Wong VK, Reuter S, Holden MTG, Fookes M, Falush
504 D, Keane JA, Parkhill J. 2015. Roary: Rapid large-scale prokaryote pan genome analysis.
505 Bioinformatics 31:3691–3693.
506 30. Seemann T. 2014. Prokka: Rapid prokaryotic genome annotation. Bioinformatics
507 30:2068–2069.
508 31. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, Bentley SD, Parkhill J, Harris
509 SR. 2015. Rapid phylogenetic analysis of large samples of recombinant bacterial whole
510 genome sequences using Gubbins. Nucleic Acids Res 43:e15.
511 32. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. 2015. IQ-TREE: A fast and
512 effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol
513 Evol 32:268–274.
514 33. Thi Hoang D, Chernomor O, von Haeseler A, Quang Minh B, Sy Vinh L, Rosenberg MS.
515 2017. UFBoot2: Improving the Ultrafast Bootstrap Approximation. Mol Biol Evol
516 35:518–522.
517 34. Kalyaanamoorthy S, Minh BQ, Wong TKF, Von Haeseler A, Jermiin LS. 2017.
518 ModelFinder: Fast model selection for accurate phylogenetic estimates. Nat Methods

519 14:587–589.
520 35. Löytynoja A. 2014. Phylogeny-aware alignment with PRANK. Methods Mol Biol
521 1079:155–170.
522 36. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden
523 DM. 2012. A program for annotating and predicting the effects of single nucleotide
524 polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118;
525 iso-2; iso-3. Fly (Austin) 6:80–92.
526 37. R Core Team. 2018. R: A language and environment for statistical computing. 3.5.0. R
527 Foundation for Statistical Computing, Vienna, Austria.
528 38. Paradis E, Schliep K. 2019. Ape 5.0: An environment for modern phylogenetics and
529 evolutionary analyses in R. Bioinformatics 35:526–528.
530 39. Yu G. 2020. aplot: Decorate a “ggplot” with Associated Information. 0.0.6.
531 40. Dowle M, Srinivasan A. 2020. data.table: Extension of ‘data.frame‘. 1.12.8.
532 41. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. Ggtree: an R Package for
533 Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other
534 Associated Data. Methods Ecol Evol 8:28–36.
535 42. Kassambara A. 2020. ggpubr: “ggplot2” Based Publication Ready Plots. 0.4.0.
536 43. Kolde R. 2019. pheatmap: Pretty Heatmaps. 1.0.12.
537 44. Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other
538 things). Methods Ecol Evol 3:217–223.
539 45. Wickham H. 2017. tidyverse: Easily Install and Load the “Tidyverse.” R package version
540 1.2.1.
541 46. 2016. Anaconda Software Distribution. 2-2.4.0.

47. Stayton CT. 2008. Is convergence surprising? An examination of the frequency of convergence in simulated datasets. J Theor Biol 252:1–14.

48. Pagel M. 1997. Inferring evolutionary processes from phylogenies. Zool Scr 26:331–348.

49. Lees JA, Galardini M, Bentley SD, Weiser JN, Corander J. 2018. pyseer: a comprehensive tool for microbial pangenome-wide association studies. Bioinformatics https://doi.org/10.1093/bioinformatics/bty539.