It is made available under a CC-BY 4.0 International license .

### 1 Title:

- 2 A multisite genomic epidemiology study of *Clostridioides difficile* infections in the U.S. supports
- 3 differential roles of healthcare versus community spread for two common strains

### 4 Authors:

- 5 Arianna Miles-Jay<sup>1</sup>, Vincent B. Young<sup>1</sup>, Eric G. Pamer<sup>2,3</sup>, Tor C. Savidge<sup>4</sup>, Mini Kamboj<sup>2</sup>, Kevin W.
- 6 Garey<sup>5</sup>, Evan S. Snitkin<sup>1#</sup>

## 7 Affiliations:

- 8 1. University of Michigan Medical School, Ann Arbor, MI, USA
- 9 2. Memorial Sloan Kettering Cancer Center, New York, NY, USA
- 10 3. Present affiliation: University of Chicago, Chicago, IL, USA
- 11 4. Baylor College of Medicine, Houston, TX, USA
- 12 5. University of Houston College of Pharmacy, Houston, TX, USA

#### 13 **# Corresponding author:**

- 14 Evan S. Snitkin
- 15 1520D MSRB I
- 16 1150 W. Medical Center Dr.
- 17 Ann Arbor, MI, 48109
- 18 Telephone: (734) 647-6472
- 19 Fax: (734) 615-5534
- 20 Email: <u>esnitkin@umich.edu</u>

It is made available under a CC-BY 4.0 International license .

- 22 Keywords: C. difficile, genomic epidemiology, transmission, healthcare, community, whole
- 23 genome sequencing
- 24 **Repositories:** All whole genome sequence data was uploaded to the National Center for
- 25 Biotechnology Information (NCBI) Sequence Read Archive (SRA) under Bioproject accessions
- 26 PRJNA595724, PRJNA561087, and PRJNA594943.
- 27
- 28

It is made available under a CC-BY 4.0 International license .

### 29 ABSTRACT

30 *Clostridioides difficile* is the leading cause of healthcare-associated infectious diarrhea.

31	However, it is increasingly appreciated that healthcare-associated infections derive from both
32	community and healthcare transmission, and that the primary sites of C. difficile transmission
33	may be strain dependent. We conducted a multisite genomic epidemiology study to assess
34	differential genomic evidence of healthcare vs. community spread for two of the most common
35	C. difficile strains in the U.S.: sequence type (ST) 1 (associated with Ribotype 027) and ST2
36	(associated with Ribotype 014/020). Isolates recovered from stool specimens collected during
37	standard clinical care at three geographically distinct U.S. medical centers between 2010 and
38	2018 underwent whole genome sequencing and phylogenetic analyses. ST1 and ST2 isolates
39	both displayed some evidence of phylogenetic clustering by study site, but clustering was
40	stronger and more apparent in ST1, consistent with our healthcare-based study more
41	comprehensively sampling local transmission of ST1 compared to ST2 strains. Analyses of
42	pairwise single nucleotide variant (SNV) distance distributions were also consistent with more
43	evidence of healthcare transmission of ST1 compared to ST2, with 44% of ST1 isolates being
44	within 2 SNVs of another isolate from the same geographic collection site compared to 5.5% of
45	ST2 isolates (p-value = <0.001). Conversely, ST2 isolates were more likely to have close genetic
46	neighbors across disparate geographic sites compared to ST1 isolates, further supporting non-
47	healthcare routes of spread for ST2 and highlighting the potential for misattributing genomic
48	similarity among ST2 isolates to recent healthcare transmission. Finally, we estimated a lower
49	evolutionary rate for the ST2 lineage compared to the ST1 lineage using Bayesian timed
50	phylogenomic analyses, and hypothesize that this may contribute to observed differences in

It is made available under a CC-BY 4.0 International license .

- 51 geographic concordance among closely related isolates. Together, these findings suggest that
- 52 ST1 and ST2, while both common causes of *C. difficile* infection in hospitals, show differential
- reliance on community and hospital spread. This conclusion supports the need for strain-
- 54 specific criteria for interpreting genomic linkages and emphasizes the importance of
- 55 considering differences in the epidemiology of circulating strains when devising interventions to
- 56 reduce the burden of *C. difficile* infections.

57

It is made available under a CC-BY 4.0 International license .

- 59 DATA SUMMARY: All whole genome sequence data was uploaded to the National Center for
- 60 Biotechnology Information (NCBI) Sequence Read Archive (SRA) under Bioproject accessions
- 61 PRJNA595724, PRJNA561087, and PRJNA594943. Metadata that comply with patient privacy
- 62 rules are included in the Supplementary Materials.
- 63
- 64

It is made available under a CC-BY 4.0 International license .

#### 65 INTRODUCTION

66	Clostridioides difficile is a gram-positive spore-forming anaerobic bacterium that is a dominant
67	cause of infectious diarrhea, colitis, and colitis-associated death in the United States [1,2].
68	While <i>C. difficile</i> infection (CDI) is classically considered nosocomial [3], recent molecular
69	epidemiologic research suggests that less than 40% of CDI cases are linkable to other
70	symptomatic CDI cases within the same hospital [4–6]. This insight has disrupted the paradigm
71	of <i>C. difficile</i> as an exclusively nosocomial pathogen and expanded interest into the roles of
72	alternative routes of <i>C. difficile</i> transmission, including community-based acquisition with
73	subsequent progression to CDI within healthcare settings [7].
74	
75	Different <i>C. difficile</i> strains may have varying propensities for transmission within healthcare vs.
76	the community, and fluroquinolone resistance has been raised as a potential defining
77	characteristic of strains that spread more readily within healthcare settings [8]. In particular,
78	the largely fluoroquinolone-resistant (FQR) Ribotype (RT) 027—also known as NAP1 via pulse-
79	field gel electrophoreses or sequence type (ST) 1 via multi-locus sequence typing (MLST)—has
80	been implicated in numerous hospital-based CDI outbreaks and is most commonly healthcare-
81	associated according to surveillance definitions based on past hospitalizations [9–12]. Another
82	common <i>C. difficile</i> lineage in the U.S., RT014/020 (corresponding to STs 2, 49, and 13), is
83	largely fluoroquinolone sensitive (FQS) and, while it is frequently characterized as healthcare-
84	associated using these same surveillance definitions, has not been associated with hospital-
85	based outbreaks [13]. Associations between C. difficile strain type and propensity for
86	healthcare-associated transmission would indicate that devising effective interventions for

It is made available under a CC-BY 4.0 International license .

87 reducing the burden of CDI may require an understanding of the molecular epidemiology of

- 88 locally circulating strains, and that strain-specific incidence may be a more accurate metric of
- 89 the successful prevention of *C. difficile* transmission within hospitals.
- 90

91	Whole genome sequencing (WGS) can provide insight into the potential contribution of
92	healthcare vs. community spread of particular strains, even in the absence of comprehensive
93	sampling of transmission networks. Recent studies that applied WGS to European clinical C.
94	difficile isolates found that RT027/ST1 displayed genomic patterns consistent with healthcare-
95	associated-spread, while RT014/020/ST2 displayed genomic patterns more consistent with
96	community-associated reservoirs [6,8]. However, these distinct epidemiologic patterns have
97	not yet been assessed using genomic data gathered from U.Sbased C. difficile isolates. Here,
98	we applied WGS to isolates collected from three geographically distinct U.S. medical centers to
99	assess differential genomic evidence of healthcare vs. community spread between two of the
100	most common <i>C. difficile</i> strains: ST1 and ST2.
101	
102	METHODS
103	Data collection
104	New C. difficile sequences were derived from clinical stool specimens collected as part existing
105	molecular surveillance programs that took place at three U.S. medical centers: Michigan
106	Medicine (UM) between 2010 and 2013 [14], Texas Medical Center Hospitals (TMC) between
107	2011 and 2017 [15], and Memorial Sloan Kettering Cancer Center (MSKCC) between 2013 and

108 2017 [16]. At all three sites, toxigenic *C. difficile* positive stool specimens were collected, *C.* 

It is made available under a CC-BY 4.0 International license .

*difficile* isolates were recovered from the speciemens, and DNA was extracted from a single
colony as previous described [14–16]. Isolates underwent molecular typing via ribotyping at UM
and TMC [17], and MLST at MSKCC [18]. DNA from a sample of isolates that were typed as
RT027 or RT014/020 at UM and TMC and ST1 or ST2 at MSKCC was extracted for whole genome
sequencing. The Institutional Review Boards at each of the study sites approved the study
protocols.

115

### 116 Whole genome sequencing and bioinformatic methods

117 DNA was sent to UM and the Nextera XT library preparation kit (Illumina, San Diego, CA) was 118 used to prepare sequencing libraries according to the manufacturer's instructions. WGS was 119 executed on an Illumina Hiseq platform with 150 base-pair paired-end reads and a targeted 120 read depth of >100X. Sequence data are available from the National Center for Biotechnology 121 Information (NCBI) Sequence Read Archive (SRA) under BioProjects PRJNA595724, 122 PRJNA561087, and PRJNA594943. The bioinformatics methods applied to the new C. difficile 123 sequences to identify single nucleotide variants (SNVs) and build phylogenetic trees were 124 executed as previously described [19]. Briefly, raw sequencing reads were trimmed using 125 Trimmomatic to remove low quality bases and adapter sequences. Trimmed reads were then 126 mapped to existing complete reference genomes within the same ST (R20291 for ST1 [GenBank 127 accession number FN545816], and W0022a for ST2 [GenBank accession number CP025046]) with the Burrows-Wheeler short read aligner [20–22]. PCR duplicates were discarded and 128 129 variants were called using SAMtools mpileup and bcftools [23]. Gubbins was used to remove 130 variant sites located in putative recombinant regions [24]. In silico multilocus sequence typing

It is made available under a CC-BY 4.0 International license .

<ul> <li>included in all analyses [25]. Maximum-likelihood phylogenies were built using IQ-TREE with a generalized time reversible nucleotide substitution model; phylogenies were rooted using <i>C</i>.</li> <li><i>difficile</i> 630 as an outgroup (GenBank accession number GCA_00009205.2)[26,27].</li> <li>Fluroquinolone resistance was assigned based off of the presence of previously identified fluroquinolone resistance-associated gryA and gyrB alleles [28]. ST1 isolates were further</li> <li>classified into previously identified FQS, FQR1, and FQR2 lineages by examining how new</li> <li>isolates clustered with publicly available FQR1 and FQR2 isolates [29].</li> </ul>
generalized time reversible nucleotide substitution model; phylogenies were rooted using <i>C</i> . <i>difficile</i> 630 as an outgroup (GenBank accession number GCA_000009205.2)[26,27]. Fluroquinolone resistance was assigned based off of the presence of previously identified fluroquinolone resistance-associated gryA and gyrB alleles [28]. ST1 isolates were further classified into previously identified FQS, FQR1, and FQR2 lineages by examining how new isolates clustered with publicly available FQR1 and FQR2 isolates [29]. <i>Evaluation of phylogenetic clustering</i> To compare the level of clustering by geographic collection site between newly sequenced ST1
<ul> <li><i>difficile</i> 630 as an outgroup (GenBank accession number GCA_000009205.2)[26,27].</li> <li>Fluroquinolone resistance was assigned based off of the presence of previously identified</li> <li>fluroquinolone resistance-associated gryA and gyrB alleles [28]. ST1 isolates were further</li> <li>classified into previously identified FQS, FQR1, and FQR2 lineages by examining how new</li> <li>isolates clustered with publicly available FQR1 and FQR2 isolates [29].</li> </ul> <b>Evaluation of phylogenetic clustering</b> To compare the level of clustering by geographic collection site between newly sequenced ST1
Fluroquinolone resistance was assigned based off of the presence of previously identified         fluroquinolone resistance-associated gryA and gyrB alleles [28]. ST1 isolates were further         classified into previously identified FQS, FQR1, and FQR2 lineages by examining how new         isolates clustered with publicly available FQR1 and FQR2 isolates [29].         Evaluation of phylogenetic clustering         To compare the level of clustering by geographic collection site between newly sequenced ST1
fluroquinolone resistance-associated gryA and gyrB alleles [28]. ST1 isolates were further classified into previously identified FQS, FQR1, and FQR2 lineages by examining how new isolates clustered with publicly available FQR1 and FQR2 isolates [29]. <i>Evaluation of phylogenetic clustering</i> To compare the level of clustering by geographic collection site between newly sequenced ST1
<ul> <li>classified into previously identified FQS, FQR1, and FQR2 lineages by examining how new</li> <li>isolates clustered with publicly available FQR1 and FQR2 isolates [29].</li> <li><i>Evaluation of phylogenetic clustering</i></li> <li>To compare the level of clustering by geographic collection site between newly sequenced ST1</li> </ul>
<ul> <li>isolates clustered with publicly available FQR1 and FQR2 isolates [29].</li> <li><i>Evaluation of phylogenetic clustering</i></li> <li>To compare the level of clustering by geographic collection site between newly sequenced ST1</li> </ul>
<i>Evaluation of phylogenetic clustering</i> To compare the level of clustering by geographic collection site between newly sequenced ST1
<b>Evaluation of phylogenetic clustering</b> To compare the level of clustering by geographic collection site between newly sequenced ST1
To compare the level of clustering by geographic collection site between newly sequenced ST1
and ST2 isolates, we overlaid geographic collection site onto the maximum-likelihood whole
genome phylogenies and applied a previously described approach for formal clustering
assessment [30]. First, we tabulated the number of isolates in a "pure" subtree of each
phylogeny—defined as a subtree made up of 2 or more isolates collected from a single
geographic site that was found in >90% of bootstrapped phylogenies. To determine whether
this number was different than would be observed by chance given the phylogenetic topology
and location frequency, we calculated an empirical p-value by randomizing geographic labels
and location frequency, we calculated an empirical p-value by randomizing geographic labels and re-calculating this number 1000 times.
and location frequency, we calculated an empirical p-value by randomizing geographic labels and re-calculating this number 1000 times.

It is made available under a CC-BY 4.0 International license .

#### 153 **Evaluation of evidence of recent transmission**

154 Evidence of recent transmission was assessed using pairwise SNV distance matrices and two analytic approaches. First, we compared the lower tail (5<sup>th</sup> percentile) of the distribution of 155 pairwise SNV distances of pairs of isolates collected from the same collection site to that same 156 metric among pairs of isolates collected from different collection sites by calculating a 5<sup>th</sup> 157 percentile SNV-distance ratio (5<sup>th</sup> percentile SNV distance within sites/5<sup>th</sup> percentile SNV 158 distance between sites). To assess whether this ratio indicated an enrichment of close linkages 159 160 within collection sites greater than could be expected by chance, we randomly permuted collection sites and re-calculated the ratio 10,000 times; an observed ratio below the 2.5% 161 162 percentile of the distribution of expected ratios was applied to support significant enrichment 163 of close genetic linkages within study sites. Second, we classified genomic linkages using an SNV-distance threshold of 2 SNVs and compared the proportion of genomically linked isolates 164 (defined as being linked to at least one other isolate) among ST1 isolates compared to those 165 166 among ST2 isolates using chi-squared tests. An SNV threshold of 2 SNVs is commonly used to 167 identify pairs of *C. difficile* isolates that were likely related via direct transmission/acquisition from a common source; this threshold is based off of evolutionary rates estimated from within-168 169 host evolution [4]. We then assessed the sensitivity of these results to larger thresholds of 5-10 170 SNVs. We also compared the proportion of isolates genomically linked to at least one isolate 171 collected from a different geographic collection site between ST1 and ST2 using chi-squared 172 tests. All analyses were completed in R v4.0.2.

173

It is made available under a CC-BY 4.0 International license .

# 175 *Estimation of evolutionary rates*

176	We applied Bayesian timed phylogenomic analyses in order to estimate and compare
177	evolutionary rates between ST1 and ST2 lineages using BEAST v1.10.4 [31]. To increase the
178	power of timed phylogenomic analyses, existing ST1 and ST2 whole genome sequences were
179	downloaded from the NCBI SRA; isolates were selected from a recent publication that compiled
180	isolates from several previous C. difficile genome collections along with their ST and sampling
181	date [32]. The combined collection of existing and new sequences was then pared down to
182	facilitate running Bayesian phylogenomic analyses. First, in an effort to maximize genetic
183	diversity, one randomly selected isolate from each pair of isolates within 2 SNVs of one another
184	was removed. Second, isolates from overrepresented geographic locations were randomly
185	downsampled until the total number of isolates was less than 425. The final list of isolates that
186	were included in these analyses can be found in Supplementary Table 1.
187	
188	We assessed the suitability of the data for timed phylogenomic analyses by examining temporal
189	signal—or the relationship between genomic differences and sampling date—using two
190	methods. First, we examined a regression of sampling time vs. root-to-tip genetic distance using
191	Tempest and BactDating [33,34]. We then formally evaluated temporal signal using date
192	randomization tests, randomly permuting the sampling dates 10 times and comparing the
193	evolutionary rate estimates and their 95% credible intervals for the random datasets to the

estimates from the real data. We report both the more relaxed and more strict criteria for

195 temporal signal assessment using this approach: with the more relaxed criteria being met if the

estimated evolutionary rate was not included in the 95% credible intervals of 10 date

It is made available under a CC-BY 4.0 International license .

randomized datasets (CR1), and the more strict being met if the 95% credible interval of the
estimated evolutionary rate did not overlap any of the 95% credible intervals of the date
randomized datasets (CR2) [35]. We proceeded with evolutionary rate estimates so long as the
data met CR1.

201

202 To select BEAST model assumptions for both the date randomization tests and the final 203 evolutionary rate estimates, we started with a general time reversible nucleotide substitution 204 model with gamma distributed rate heterogeneity and the simplest clock and demographic 205 model assumptions: a strict molecular clock and constant demographic prior. We then 206 systematically examined to what extent the data violated the strict clock and constant 207 demographic model prior assumptions and thus, to what extent more complex models were warranted. To assess whether the data violated a strict clock assumption, we evaluated 208 209 whether the coefficient of variation parameter in the models with an uncorrelated relaxed 210 lognormal clock had a 95% highest posterior density interval (HPD) that overlapped 0; if not, we 211 used this as evidence of the assumptions of a strict clock being violated and applied an 212 uncorrelated relaxed lognormal clock with a lognormal prior distribution with a mean of 5.0x10<sup>-</sup> <sup>7</sup> and standard deviation of 8x10<sup>-7</sup> based on previous evolutionary rate estimates (while 213 214 allowing still allowing for significant deviation) [29,36]. To assess to what extent the data 215 violated a constant demographic model, we ran models with exponential growth demographic 216 model prior, and evaluated whether the 95% credible interval of the exponential growth rate 217 parameter overlapped 0. If the exponential growth rate parameter was substantially different 218 from 0, we attempted running a more flexible but parameter rich Gaussian Markov Random

It is made available under a CC-BY 4.0 International license .

Field (GMRF) skyride model, which allows for periods of growth as well as periods of stasis [37]. For each model, a Markov-chain Monte Carlo was run for 200 million generations and sampled every 10,000 iterations; a Tempest-rooted starting tree was included in all runs to accelerate convergence [33]. All ESS values were checked for being above 200 using Tracer after removing the first 10% of steps as burn-in [38].

224

225 **RESULTS** 

226 There were 382 new whole genome sequences generated from the 3 U.S. study sites located in 227 Michigan, Texas, and New York; 199 ST1 and 183 ST2 (Supplementary Figure 1). The majority of 228 ST1 isolates were FQR, relatively evenly distributed between the previously described FQR1 and 229 FQR2 lineages, and the FQS isolates clustered together in one ancestral clade. Conversely, ST2 230 isolates were largely FQS, with FQR isolates occurring in a two small clusters as well as 231 singletons scattered throughout the phylogeny (Figure 1). ST1 sequences were less diverse than 232 ST2 sequences: after quality and recombination filtering, the ST1 alignment consisted of 1108 233 SNVs (median pairwise SNV distance 35, range 0-85), while the ST2 alignment consisted of 2119 234 SNVs (median pairwise SNV distance 52, range 1-156) (Figure 2).

It is made available under a CC-BY 4.0 International license .

- **Figure 1:** Maximum likelihood phylogenetic trees of newly sequenced *C. difficile* isolates that
- are ST1 and ST2. Tips are colored by fluroquinolone-resistant (FQR) vs. fluroquinolone-sensitive
- 237 (FQS) as determined by the presence of previously identified fluroquinolone-resistance-
- associated gryA and gyrB alleles. Previously identified ST1 lineages (FQS, FQR1, and FQR2) are
- highlighted, collection site is included in an adjacent heatmap. Tree scales are in single
- 240 nucleotide changes per quality- and recombination-filtered site.



It is made available under a CC-BY 4.0 International license .

### 242 ST1 exhibits stronger evidence of phylogenetic clustering by geography compared to ST2

243 To begin our comparison of ST1 and ST2 isolates, we first examined the association between 244 phylogenetic and geographic structure by overlaying the geographic site each isolate was 245 collected from onto strain-specific whole genome phylogenies. Visual examination of these 246 phylogenies revealed a striking difference in geographic clustering, with ST1 displaying larger 247 clusters and ST2 displaying more numerous, smaller clusters and more geographic mixing 248 (Figure 1). The exception to this observation was the FQS ST1 clade, which appeared more 249 geographically mixed than the FQR ST1 clades. While statistical assessments demonstrated that 250 both ST1 and ST2 displayed more evidence of geographic clustering than would be expected to 251 occur by chance (empiric p-values both <0.001), clustering was more non-random for ST1 than 252 ST2 (Supplementary Figure 2). This enhanced geographic clustering among ST1 isolates could 253 reflect that our healthcare-based study more completely sampled local transmission networks 254 among ST1 isolates compared to ST2 isolates, or it could reflect ST1 spreading via more 255 localized community or healthcare reservoirs with minimal long-distance transmission.

256

257 **ST1** isolates display more evidence of recent transmission than **ST2**, while **ST2** isolates are 258 more likely to share intermediate genetic linkages across disparate geographic sites

To further investigate whether plausible healthcare-associated transmission among ST1 isolates was driving the geographic clustering patterns we saw in the phylogenies, we next examined the prevalence and nature of close genetic linkages within each lineage as captured by pairwise SNV distances. Isolates linked by very small SNV distances are plausibly linked via recent transmission, and we would expect our healthcare-based study to more comprehensively

It is made available under a CC-BY 4.0 International license .

264	sample healthcare-associated transmission than community-associated transmission. When
265	examining the SNV distance distributions between and within collection sites, among ST1
266	isolates, we observed more closely related pairs of isolates from the same geographic collection
267	site (reflected by a heavier lower tail of the distribution) compared to pairs of isolates collected
268	from different geographic collection sites (5 <sup>th</sup> percentile SNV distance within sites/5 <sup>th</sup> percentile
269	SNV distance between sites = 0.59, expected ratio 95% interval 0.93-1.00, Figure 2). However,
270	we did not observe this same pattern among ST2 isolates (5 <sup>th</sup> percentile SNV distance within
271	sites/5 <sup>th</sup> percentile SNV distance between sites = 1.00, expected ratio 95% interval 0.93-1.00,
272	Figure 2). Application of SNV distance thresholds demonstrated that 88 (44%) ST1 isolates were
273	within 2 SNVs of another isolate from the same geographic collection site compared to 10
274	(5.5%) ST2 isolates (p-value = <0.001). As the SNV threshold was increased to intermediate
275	values of 5 and 10 SNVs, this trend was maintained (all p < 0.001, Figure 3A). Conversely, at the
276	5 and 10 SNV thresholds, linked ST2 isolates were more likely to be linked to an isolate from a
277	different geographic collection site compared to linked ST1 isolates (all p < 0.001, Figure 3A).
278	These geographically discordant intermediate genomic linkages among ST2 were not associated
279	with temporal linkages, with the days between sample collection ranging from 6 to 2,479 days
280	(Figure 3B). Among geographically discordant ST1 isolates pairs, FQS isolates were
281	overrepresented, with the only pair of geographically discordant ST1 isolates linked within 5
282	SNVs being FQS and 14/31 (45.2%) geographically discordant ST1 isolates linked within 10 SNVs
283	being FQS even though FQS isolates made up only 21/199 (10.6%) of isolates overall. Together,
284	these findings are consistent with evidence of recent healthcare transmission among ST1
285	isolates and transmission outside of the hospital among ST2 isolates, and also raise questions

It is made available under a CC-BY 4.0 International license .

- about the underlying reasons why ST2 isolates are more likely to be closely related across
- 287 disparate geographic sites.

288

- 289 Figure 2: Pairwise single nucleotide variant (SNV) distribution between pairs of isolates from
- the same collection site vs. pairs of isolates from geographically distinct collection sites for both
- 291 ST1 and ST2. The black diamond indicates fifth percentile SNV distances for each category.
- 292



It is made available under a CC-BY 4.0 International license .

Figure 3: A) Bar plot showing the proportion of ST1 and ST2 isolates that are genomically linked to another isolate, either from the same collection site (green) only or from at least one different collection site (orange), at varying SNV thresholds. B) Scatter plot of days between collection and pairwise SNV distance up to 10 SNVs, where each dot represents one pair of isolates. Points are colored by whether they are collected from the same geographic collection site (green) or different geographic collection sites (orange). Points are jittered to improve clarity.



It is made available under a CC-BY 4.0 International license .

#### 302 Timed phylogenomic analyses demonstrate evidence of evolutionary rate heterogeneity

### 303 within and between ST1 and ST2 lineages

304 Our observation that ST2 is more likely to be genomically linked at intermediate SNV thresholds 305 across disparate geographic sites compared to ST1 isolates led us to explore the potential 306 mechanisms underlying this difference. Two factors we hypothesized might contribute to these 307 findings are 1) increased transmission of ST2 via community-based reservoirs that facilitate 308 more rapid spread over large geographic distances and/or 2) a slower average evolutionary rate 309 among ST2 isolates resulting in less genetic changes over larger amounts of time and space. 310 While examining the former hypothesis was beyond the scope of this study, we explored the 311 plausibility of the latter hypothesis by estimating evolutionary rates for ST1 and ST2 using the 312 BEAST Bayesian phylogenetic software [31]. There were 418 ST1 and 418 ST2 isolates included 313 in this analysis; sequences included a mix of newly sequenced and publicly available global 314 genomes in order to maximize temporal and genetic diversity while maintaining a sample size 315 manageable by the BEAST software (Supplementary Table 1). For ST1 isolate selection, we also 316 opted to maintain all FQS ST1 isolates, given our observations that they may display distinct 317 epidemiological patterns from FQR ST1 isolates.

318

Temporal signal analyses, while initiated as a necessary precursor to timed phylogenomic analyses in BEAST, revealed interesting differences between the clock-like nature of ST1 and ST2 isolates. While root-to-tip regression analyses suggested similarly weak but sufficient temporal signal to proceed with timed phylogenomic analyses in BEAST (indicated by positive correlation coefficients, Supplementary Figure 3), the more rigorous hypothesis testing date

It is made available under a CC-BY 4.0 International license .

324	randomization tests demonstrated more evidence of temporal signal among ST1 isolates, which
325	passed both the more relaxed CR1 and more stringent CR2 criteria for temporal analyses,
326	compared to ST2 isolates, which passed CR1 but not CR2 (Supplementary Figure 4). The root-
327	to-tip regression also highlighted different temporal patterns among FQS-ST1 isolates
328	compared to FQR-ST1 isolates, which was observed again in date randomization tests on FQS-
329	ST1 and FQR-ST1 isolates separately; the FQR-ST1 isolates appeared to drive the temporal
330	signal in the data, and when considered alone, FQS-isolates were more like ST2 isolates, passing
331	the more relaxed CR1 temporal signal criteria but not the more stringent CR2. This observation
332	was consistent with our pairwise SNV distance findings of distinct patterns among FQS ST1
333	isolates, and motivated conducting further analyses both with all ST1 isolates together as well
334	as with FQR ST1 isolates (n = 359) and FQS ST1 isolates (n = 59) considered separately.
335	
336	All datasets demonstrated evidence of evolutionary rate heterogeneity throughout the
337	phylogeny, resulting in the application of uncorrelated relaxed lognormal molecular clock

338 models along with a constant demographic priors (see Supplementary Figures 5-6 and 339 Supplementary Results for details). Overall, when considering all ST1 isolates together 340 compared to all ST2 isolates, evolutionary rate estimates were slightly higher for ST1 compared 341 to ST2, although the 95% credible intervals overlapped. However, ST1's faster evolutionary rate 342 was driven by FQR ST1 isolates; when separating out FQS and FQR ST1 isolates, the FQR ST1 evolutionary rate estimates emerged as significantly higher than that of ST2 isolates (with non-343 344 overlapping 95% credible intervals) while FQS ST1 isolates had similar evolutionary rate 345 estimates to ST2 isolates (Figure 4). These evolutionary rates translate to approximately 1.36

346	(95% credible interval 1.20-1.52) nucleotide changes per year for FQR ST1, 0.80 (95% credible
347	interval 0.51-1.08) nucleotide changes per year for FQS-ST1, and 0.89 (95% credible interval
348	0.74-1.05) nucleotide changes per year for ST2. These results are consistent with the hypothesis
349	that a slightly slower average evolutionary rate among ST2 and FQS ST1 isolates compared to
350	FQR ST1 isolates might contribute to our observed discordance between genomic and
351	epidemiologic linkages among those isolates.
352	
353	Figure 4: Posterior probability density of the evolutionary rates estimates for C. difficile ST1 and
354	ST2 lineages, with ST1 isolates considered together as well as separated out into FQR-ST1 and
355	FQS-ST1 isolates. Dark shaded areas of the density curves indicate the lower 2.5% and upper
356	97.5% of the distributions; light shaded areas indicate 95% credible intervals. Evolutionary rates

357 are considered significantly different from one another when the 95% credible intervals of their





360

It is made available under a CC-BY 4.0 International license .

#### 361 **DISCUSSION**

362	In this study, we investigated the genomic epidemiology of two dominant C. difficile lineages,
363	ST1 and ST2, across three geographically distinct U.S. medical centers. We observed more
364	genomic evidence of geographic clustering and recent transmission among ST1 isolates
365	compared to ST2 isolates, while also finding more linkages among ST2 isolates from disparate
366	geographic collection sites at intermediate genomic linkage thresholds. Lastly, we estimated a
367	slightly more rapid average evolutionary rate for FQR ST1 isolates compared to FQS ST1 isolates
368	and ST2 isolates using Bayesian timed phylogenomic methods.
369	

370 Previous studies have reported both more evidence of broad geographic clustering [8] and 371 more evidence of recent transmission within healthcare settings [6] among European ST1 C. 372 *difficile* isolates compared to other types of *C. difficile*. To our knowledge, these are the first 373 U.S.-based multisite data to support these findings. Our observations are consistent with ST1 374 being associated with hospital outbreaks [9–11], being the most predominant healthcare-375 associated C. difficile strain according to surveillance definitions based on timing since last 376 healthcare exposure [13], and being more prevalent in hospital than community environmental 377 sampling [39]. The factors contributing to increased spread of ST1 within healthcare are not 378 well defined, however, fluroquinolone resistance has been proposed as a driving feature. In 379 support of this, Eyre et al. noted that other FQR C. difficile strains were also more likely to cluster by country compared to FQS C. difficile strains [13]. Our observations of distinct 380 381 epidemiological and evolutionary patterns among FQS compared to FQR ST1 isolates are also 382 consistent with this hypothesis. If within-healthcare transmission is the dominant mode of ST1

It is made available under a CC-BY 4.0 International license .

spread, infection control interventions and antimicrobial-stewardship within healthcare should
jointly reduce the incidence of CDI due to ST1. Such reductions have been reported in the UK
after implementation of national infection prevention and antimicrobial stewardship policies
[40].
Conversely, ST2 seems to have followed a different route to pathogenic success. RT014/ST2 has

389 been reported as one of the most common strains In Europe [41], the US [13,42,42], and 390 Australia [43] during the last decade. ST2 is commonly characterized in the literature as an 391 endemic strain in the U.S. that has not been associated with hospital outbreaks [44]. However, 392 it is also frequently classified as healthcare associated: the most recent data from the Centers 393 for Disease Control and Prevention Emerging Infections Program reports between 41% and 52% of RT014 were considered healthcare-associated infections between 2012 and 2017 [13]. 394 Despite this, evidence of transmission of RT014/ST2 within the hospital is sparse, as 395 396 demonstrated by this study and others [6,13]. One explanation for this discordance between 397 genomic evidence of recent transmission and healthcare-associated characterization via 398 surveillance definitions is that ST2 is frequently acquired in the community, imported into the 399 hospital, and subsequently progresses to infection after hospitalization. If this is the case, 400 antimicrobial stewardship interventions may be particularly effective for preventing infections 401 due to this common strain [7]. Environmental studies that have reported recovery of RT014 402 isolates in agriculture [45,46], wastewater [47], and parks and homes [39] are also consistent 403 with community circulation of RT014. Overall, this finding highlights the imperfect nature of relying on infection onset as a proxy for acquisition. With the advent of more widespread 404

It is made available under a CC-BY 4.0 International license .

pathogen whole genome sequencing, genomic evidence of healthcare transmission could be
used as an alternative and more accurate metric than infection onset for measuring withinhospital transmission of *C. difficile*.
We also observed a notable difference in concordance between genomic linkages (isolate

410 related within small SNV distance thresholds) and epidemiologic linkages (isolates collected 411 from the same site within temporally proximate time periods) among ST1 and ST2 isolates. 412 Specifically, ST2 isolates were more likely to have close genomic neighbors across disparate 413 geographic sites and long time periods. Consistent with this, a pan-European surveillance study 414 reported that the average most closely related strain to any given RT014 isolate was collected 415 from hundreds of miles away [13]. The mechanisms behind this finding are not clear, but are 416 consistent with a reliance on non-healthcare routes of spread. Practically speaking, this finding highlights the risks of broadly applying SNV thresholds to infer recent transmission, even to 417 418 isolates of the same species. In particular, it emphasizes the importance of considering 419 background genomic diversity and incorporating geographically and temporally diverse strains 420 when interpreting genomic linkages. Without this context, one might mistakenly attribute a 421 linkage to transmission when it in fact reflects broader genomic diversity patterns in a 422 particular lineage. The importance of genomic context has been noted since the early days of 423 bacterial genomic epidemiology [48], but in most cases, sequencing is still not widespread 424 enough to provide such context. As we continue to consider a future with routine genomic 425 surveillance in hospital settings to identify outbreaks [49], it is crucial that assessment of

It is made available under a CC-BY 4.0 International license .

genomic context remain part of the evidence required for inferring transmission from genomicdata.

428

429 *C. difficile*'s spore-forming lifestyle may contribute to some of the results reported here. It has 430 been posited that spore formation likely drags down average estimate evolutionary rates of 431 bacteria [50]. Extending from that, if isolates belonging to particular lineages spend more time 432 in spore form than others, that lineage could be expected to have a lower average evolutionary 433 rate, and thus, less nucleotide differences accumulated over time. We speculate that the ST2 and FQS ST1 lineages may have spent, on average, more time in spore-form than the epidemic 434 435 and more recently emerged FQR ST1 lineages resulting in more closely related isolates across 436 larger amounts of time and space. Ecological niches may influence this: more selective pressures and a higher density of susceptible hosts in healthcare settings could facilitate more 437 time in the vegetative state, whereas strains that circulate primarily in the community may be 438 439 more likely to stay dormant for longer periods of time. Results from our Bayesian timed phylogenomic analyses were consistent with this framework in two ways: 1) high evolutionary 440 441 rate heterogeneity in both ST1 and ST2 isolates may reflect the effects of spore formation, with 442 isolates emerging for a long-dormant spore being found on the tips of phylogenetic branches with a slow estimated evolutionary rate and 2) less evidence of temporal signal and slightly 443 444 lower estimated evolutionary rates for FQS-ST1 isolates and ST2 isolates compared to FQR-ST1 445 isolates may reflect more time spent in spore-form. Whatever the biological and 446 epidemiological underpinnings of the patterns we observed, this work highlights the challenges

It is made available under a CC-BY 4.0 International license .

inherent to applying molecular clock-based methods to studying the epidemiology and

448 evolution of a variably and relatively slowly evolving pathogen like *C. difficile*.

449

Our findings should be interpreted in the context of multiple limitations. First, the retrospective 450 451 nature of the study resulted in some differences in sample collection between the three study 452 sites: UM and TMC selected based off of PCR Ribotypes, which we then filtered down to only 453 ST1 and ST2 via in silico MLST, while MSKCC originally selected isolates based off of ST as MLST 454 is routine at that center. However, all comparisons were made between ST1 and ST2 isolates 455 and these differences were consistent within the ST1 and ST2 isolates at each site, so we would 456 not expect them to significantly alter the results reported here. Second, limited epidemiologic 457 metadata was available for analysis: only study site and collection date. Despite this, the 458 interesting patterns we observed between genomic linkages and epidemiologic linkages emphasizes the value of integrating genomic data with even limited epidemiologic metadata. 459 460 Finally, the evolutionary rate estimates presented here are subject to uncertainty, particularly given the observed instances of violated model assumptions and relatively limited temporal 461 462 signal in the data. However, the overall trends remained stable with varying models, alleviating 463 concerns that our findings are artifacts of model misspecification. This study also has several notable strengths, including the collection of isolates from three distinct geographic sites in the 464 465 U.S., the application of whole genome sequencing for high-resolution typing and phylogenetic 466 analyses, and the incorporation of global isolates for increased context and power in our timed 467 phylogenomic analyses.

It is made available under a CC-BY 4.0 International license .

# 469 **Conclusions**

470	Examination of the genomic epidemiology of <i>C. difficile</i> ST1 and ST2 across three geographically
471	distinct U.S. medical centers revealed divergent epidemiologic and evolutionary patterns
472	between these two common strains. Specifically, we observed more evidence of geographic
473	clustering, recent healthcare transmission, and a slightly more rapid average evolutionary rate
474	among FQR ST1 isolates compared to ST2 and FQS ST1 isolates. One implication of these
475	findings is that an understanding of local molecular epidemiology may facilitate the
476	development of effective interventions targeted at reducing the burden of CDI. These findings
477	also highlight how methodological considerations—including incorporating genomic context
478	when inferring transmission from genomic linkages and considering the potential effect of
479	spore formation on the connection between genomic differences and epidemiology—need to
480	be accounted for when applying genomic epidemiology methods for studying C. difficile
481	transmission.

It is made available under a CC-BY 4.0 International license .

### 483 **ACKNOWLEDGEMENTS**

484 We thank Ali Pirani for his contributions to the bioinformatics data analyses.

It is made available under a CC-BY 4.0 International license .

# 485 FUNDING

- 486 This work was supported by the National Institutes of Health via U01AI12455 (A.M-J., E.S.,
- 487 V.B.Y.), the Molecular Mechanisms of Microbial Pathogenesis Training Grant (T32AI007528,
- 488 A.M-J.), U01AI124290 (T.C.S., K.W.G.), and U01AI124275 (E.G.P., M.K.).
- 489

It is made available under a CC-BY 4.0 International license .

# 490 CONFLICTS OF INTEREST

491 The authors declare that there are no conflicts of interest.

It is made available under a CC-BY 4.0 International license .

### 492 **REFERENCES**

- Magill SS, O'Leary E, Janelle SJ, et al. Changes in Prevalence of Health Care–Associated
   Infections in U.S. Hospitals. N Engl J Med. 2018; 379(18):1732–1744.
- Lessa FC, Mu Y, Bamberg WM, et al. Burden of iClostridium difficile/i Infection in the
  United States. N Engl J Med. **2015**; 372(9):825–834.
- 497 3. Martin JSH, Monaghan TM, Wilcox MH. Clostridium difficile infection: Epidemiology,
  498 diagnosis and understanding transmission. Nat Rev Gastroenterol Hepatol. 2016;
  499 13(4):206–216.
- 500 4. Eyre DW, Cule ML, Wilson DJ, et al. Diverse sources of C. difficile infection identified on 501 whole-genome sequencing. N Engl J Med. **2013**; 369(13):1195–205.
- Walker AS, Eyre DW, Wyllie DH, et al. Characterisation of clostridium difficile hospital
   ward-based transmission using extensive epidemiological data and molecular typing. PLoS
   Med. 2012; 9(2):e100172.
- Martin JSH, Eyre DW, Fawley WN, et al. Patient and strain characteristics associated with
   Clostridium difficile transmission and adverse outcomes. Clin Infect Dis. **2018**; 67(9):1379 1387.
- Poirier D, Gervais P, Fuchs M, et al. Predictors of Clostridioides difficile Infection Among
   Asymptomatic, Colonized Patients: A Retrospective Cohort Study. Clin Infect Dis. 2020;
   70(10):2103-2210.
- Eyre DW, Davies KA, Davis G, et al. Two Distinct Patterns of Clostridium difficile Diversity
   Across Europe Indicating Contrasting Routes of Spread. Clin Infect Dis. 2018; 67(7):1035–
   1044.
- Loo VG, Oughton M, Bourgault A-M, Kelly M, Dewar K, Monczak Y. A Predominantly Clonal
   Multi-Institutional Outbreak of Clostridium difficile–Associated Diarrhea with High
   Morbidity and Mortality. N Engl J Med. 2005; 353(23):2442-2449.
- 517 10. McDonald LC, Owens RC, Johnson S. An Epidemic, Toxin Gene–Variant Strain of
  518 Clostridium difficile. N Engl J Med. 2005; 353(23):2433-2441.
- Warny M, Pepin J, Fang A, et al. Toxin production by an emerging strain of Clostridium
   difficile associated with outbreaks of severe disease in North America and Europe. 2005;
   366:6.
- Kuijper EJ, Barbut F, Brazier JS, et al. Update of Clostridium difficile infection due to PCR
   ribotype 027 in Europe, 2008. Euro Surveill. **2008**; 13(31):pii=18942.
- 524

- 525 13. Guh AY, Mu Y, Winston LG, et al. Trends in U.S. Burden of *Clostridioides difficile* Infection
  526 and Outcomes. N Engl J Med. **2020**; 382(14):1320–1330.
- Rao K, Micic D, Natarajan M, et al. Clostridium difficile Ribotype 027: Relationship to Age,
   Detectability of Toxins A or B in Stool with Rapid Testing, Severe Infection, and Mortality.
   Clin Infect Dis. **2015**; 61(2):233–241.
- 530 15. Gonzales-Luna AJ, Carlson TJ, Dotson KM, et al. PCR ribotypes of *Clostridioides difficile*531 across Texas from 2011 to 2018 including emergence of ribotype 255. Emerg Microbes
  532 Infect. 2020; 9(1):341–347.
- Kamboj M, McMillen T, Syed M, et al. Evaluation of a combined Multi-Locus Sequence
   Typing and Whole-Genome Sequencing Two-step Algorithm for Routine typing of
   *Clostridioides difficile*. J Clin Microbiol. **2020**; JCM.01955-20:*Online ahead of print*.
- 536 17. Martinson JNV, Broadaway S, Lohman E, et al. Evaluation of Portability and Cost of a
  537 Fluorescent PCR Ribotyping Protocol for Clostridium difficile Epidemiology. Onderdonk AB,
  538 editor. J Clin Microbiol. **2015**; 53(4):1192–1197.
- 539 18. Griffiths D, Fawley W, Kachrimanidou M, et al. Multilocus Sequence Typing of *Clostridium* 540 *difficile*. J Clin Microbiol. **2010**; 48(3):770–778.
- 541 19. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence
  542 data. Bioinformatics. 2014; 30(15):2114–2120.
- 543 20. He M, Sebaihia M, Lawley TD, et al. Evolutionary dynamics of Clostridium difficile over
  544 short and long time scales. Proc Natl Acad Sci. 2010; 107(16):7527–7532.
- 545 21. Yin C, Chen DS, Zhuge J, et al. Complete Genome Sequences of Four Toxigenic *Clostridium*546 *difficile* Clinical Isolates from Patients of the Lower Hudson Valley, New York, USA.
  547 Genome Announc. **2018**; 6(4):e01537-17.
- 548 22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform.
  549 Bioinformatics. 2009; 25(14):1754–1760.
- Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools.
  Bioinformatics. **2009**; 25(16):2078–2079.
- 552 24. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of
  553 recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. 2015;
  554 43(3):e15.
- 555 25. Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: rapid antimicrobial resistance
  556 genotyping directly from sequencing reads. Microb Genomics. 2017; 3(10):e000131.

It is made available under a CC-BY 4.0 International license .

- S57 26. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic
  algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. 2015; 32(1):268–
  274.
- Sebaihia M, Wren BW, Mullany P, et al. The multidrug-resistant human pathogen
  Clostridium difficile has a highly mobile, mosaic genome. Nat Genet. **2006**; 38(7):779–786.
- 562 28. Spigaglia P, Barbanti F, Mastrantonio P, Brazier JS. Fluoroquinolone resistance in
  563 Clostridium difficile isolates from a prospective study of C. difficile infections in Europe. J
  564 Med Microbiol. 2008; 57:744-789.
- He M, Miyajima F, Roberts P, et al. Emergence and global spread of epidemic healthcareassociated Clostridium difficile. Nat Genet. **2013**; 45(1):109–113.
- 30. Popovich KJ, Snitkin ES, Hota B, et al. Genomic and Epidemiological Evidence for
  Community Origins of Hospital-Onset Methicillin-Resistant Staphylococcus aureus
  Bloodstream Infections. J Infect Dis. 2017; 215:1640–1647.
- Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian
  phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. 2018;
  4(1):vey016.
- 573 32. Eyre DW, Didelot X, Buckley AM, et al. Clostridium difficile trehalose metabolism variants
  574 are common and not associated with adverse patient outcomes when variably present in
  575 the same lineage. EBioMedicine. **2019**;43:347-355.
- 33. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of
  heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. 2016;
  2(1):vew007.
- 579 34. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. Bayesian inference of ancestral
  580 dates on bacterial phylogenetic trees. Nucleic Acids Res. **2018**; 46(22):e134–e134.
- 581 35. Duchêne S, Duchêne D, Holmes EC, Ho SYW. The Performance of the Date-Randomization
  582 Test in Phylogenetic Analyses of Time-Structured Virus Data. Mol Biol Evol. 2015;
  583 32(7):1895–1906.
- 58436.Didelot X, Eyre DW, Cule M, et al. Microevolutionary analysis of Clostridium difficile585genomes to investigate transmission. Genome Biol. **2012**; 13(12):R118.
- 37. Minin VN, Bloomquist EW, Suchard MA. Smooth Skyride through a Rough Skyline:
  Bayesian Coalescent-Based Inference of Population Dynamics. Mol Biol Evol. 2008;
  25(7):1459–1471.
- 38. Rambaut A, Suchard M, Xie D, Drummond A. Tracer v1.6. 2014; . Available from:
  http://beast.bio.ed.ac.uk/Tracer

- 39. Alam MJ, Walk ST, Endres BT, et al. Community Environmental Contamination of Toxigenic
  Clostridium difficile. Open Forum Infect Dis. 2017; 4(1):ofx018.
- 593 40. Dingle KE, Didelot X, Quan TP, et al. Effects of control interventions on Clostridium difficile
  594 infection in England: an observational study. Lancet Infect Dis. 2017; 17(4):411–421.
- 595 41. Davies KA, Ashwin H, Longshaw CM, Burns DA, Davis GL, Wilcox MH. Diversity of
  596 Clostridium difficile PCR ribotypes in Europe: results from the European, multicentre,
  597 prospective, biannual, point-prevalence study of Clostridium difficile infection in
  598 hospitalised patients with diarrhoea (EUCLID), 2012 and 2013. Euro Surveill Bull Euro
  599 Surveill. 2016; 21(29):pii=30294.
- 42. Tenover FC, Tickler IA, Persing DH. Antimicrobial-Resistant Strains of Clostridium difficile
  from North America. Antimicrob Agents Chemother. **2012**; 56(6):2929–2932.
- Foster NF, Collins DA, Ditchburn SL, et al. Epidemiology of Clostridium difficile infection in
  two tertiary-care hospitals in Perth, Western Australia: a cross-sectional study. New
  Microbes New Infect. 2014; 2(3):64–71.
- Aitken SL, Alam MJ, Khaleduzzuman M, et al. In the Endemic Setting, *Clostridium difficile*Ribotype 027 Is Virulent But Not Hypervirulent. Infect Control Hosp Epidemiol. **2015**;
  36(11):1318–1323.
- Knight DR, Squire MM, Collins DA, Riley TV. Genome Analysis of Clostridium difficile PCR
  Ribotype 014 Lineage in Australian Pigs and Humans Reveals a Diverse Genetic Repertoire
  and Signatures of Long-Range Interspecies Transmission. Front Microbiol. 2017; 7:2138
- 46. Janezic S, Zidaric V, Pardon B, et al. International Clostridium difficile animal strain
  collection and large diversity of animal associated strains. BMC Microbiol. 2014; 14(1):173.
- 613 47. Romano V, Pasquale V, Krovacek K, Mauri F, Demarta A, Dumontet S. Toxigenic
  614 Clostridium difficile PCR Ribotypes from Wastewater Treatment Plants in Southern
  615 Switzerland. Appl Environ Microbiol. **2012**; 78(18):6643–6646.
- 616 48. Croucher NJ, Harris SR, Grad YH, Hanage WP. Bacterial genomes in epidemiology–present
  617 and future. Philos Trans R Soc Lond B Biol Sci. **2013**; 368(1614):20120202.
- 49. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection
  by leading with genomic surveillance of nosocomial pathogens. Microbiology. 2018;
  164(10):1213–1219.
- 50. Weller C, Wu M. A generation-time effect on the rate of molecular evolution in bacteria.
  Evolution. 2015; 69(3):643–652.
- 623