Title:

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

Authors:

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

Affiliations:

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

 # Corresponding author:

Evan S. Snitkin

1520D MSRB I

1150 W. Medical Center Dr.

Ann Arbor, MI, 48109

Telephone: (734) 647-6472

Fax: (734) 615-5534

Email: esnitkin@umich.edu

Keywords: *C. difficile,* genomic epidemiology, transmission, healthcare, community, whole genome sequencing

Repositories: All whole genome sequence data was uploaded to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under Bioproject accessions PRJNA595724, PRJNA561087, and PRJNA594943.

ABSTRACT

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

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
53 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
58

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

INTRODUCTION

Clostridioides difficile is a gram-positive spore-forming anaerobic bacterium that is a dominant cause of infectious diarrhea, colitis, and colitis-associated death in the United States [1,2]. While *C. difficile* infection (CDI) is classically considered nosocomial [3], recent molecular epidemiologic research suggests that less than 40% of CDI cases are linkable to other symptomatic CDI cases within the same hospital [4–6]. This insight has disrupted the paradigm of *C. difficile* as an exclusively nosocomial pathogen and expanded interest into the roles of alternative routes of *C. difficile* transmission, including community-based acquisition with subsequent progression to CDI within healthcare settings [7].

Different *C. difficile* strains may have varying propensities for transmission within healthcare vs. the community, and fluroquinolone resistance has been raised as a potential defining characteristic of strains that spread more readily within healthcare settings [8]. In particular, the largely fluoroquinolone-resistant (FQR) Ribotype (RT) 027—also known as NAP1 via pulse-field gel electrophoreses or sequence type (ST) 1 via multi-locus sequence typing (MLST)—has been implicated in numerous hospital-based CDI outbreaks and is most commonly healthcare-associated according to surveillance definitions based on past hospitalizations [9–12]. Another common *C. difficile* lineage in the U.S., RT014/020 (corresponding to STs 2, 49, and 13), is largely fluoroquinolone sensitive (FQS) and, while it is frequently characterized as healthcare-associated using these same surveillance definitions, has not been associated with hospital-based outbreaks [13]. Associations between *C. difficile* strain type and propensity for healthcare-associated transmission would indicate that devising effective interventions for

reducing the burden of CDI may require an understanding of the molecular epidemiology of locally circulating strains, and that strain-specific incidence may be a more accurate metric of the successful prevention of *C. difficile* transmission within hospitals.

Whole genome sequencing (WGS) can provide insight into the potential contribution of healthcare vs. community spread of particular strains, even in the absence of comprehensive sampling of transmission networks. Recent studies that applied WGS to European clinical *C. difficile* isolates found that RT027/ST1 displayed genomic patterns consistent with healthcare-associated-spread, while RT014/020/ST2 displayed genomic patterns more consistent with community-associated reservoirs [6,8]. However, these distinct epidemiologic patterns have not yet been assessed using genomic data gathered from U.S.-based *C. difficile* isolates. Here, we applied WGS to isolates collected from three geographically distinct U.S. medical centers to assess differential genomic evidence of healthcare vs. community spread between two of the most common *C. difficile* strains: ST1 and ST2.

METHODS

Data collection

New *C. difficile* sequences were derived from clinical stool specimens collected as part existing molecular surveillance programs that took place at three U.S. medical centers: Michigan Medicine (UM) between 2010 and 2013 [14], Texas Medical Center Hospitals (TMC) between 2011 and 2017 [15], and Memorial Sloan Kettering Cancer Center (MSKCC) between 2013 and 2017 [16]. At all three sites, toxigenic *C. difficile* positive stool specimens were collected, *C.*

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.

Whole genome sequencing and bioinformatic methods

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

(MLST) was performed using ARIBA and only isolates that were identified as ST1 and ST2 were 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 *C. difficile* 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].

Evaluation of phylogenetic clustering

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 re-calculating this number 1000 times.

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

154 Evidence of recent transmission was assessed using pairwise SNV distance matrices and two
155 analytic approaches. First, we compared the lower tail (5^{th} percentile) of the distribution of
156 pairwise SNV distances of pairs of isolates collected from the same collection site to that same
157 metric among pairs of isolates collected from different collection sites by calculating a 5^{th}
158 percentile SNV-distance ratio (5^{th} percentile SNV distance within sites/5^{th} percentile SNV
159 distance between sites). To assess whether this ratio indicated an enrichment of close linkages
160 within collection sites greater than could be expected by chance, we randomly permuted
161 collection sites and re-calculated the ratio 10,000 times; an observed ratio below the 2.5%
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
164 SNV-distance threshold of 2 SNVs and compared the proportion of genomically linked isolates
165 (defined as being linked to at least one other isolate) among ST1 isolates compared to those
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
168 from a common source; this threshold is based off of evolutionary rates estimated from within-
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

174

Estimation of evolutionary rates

We applied Bayesian timed phylogenomic analyses in order to estimate and compare evolutionary rates between ST1 and ST2 lineages using BEAST v1.10.4 [31]. To increase the power of timed phylogenomic analyses, existing ST1 and ST2 whole genome sequences were downloaded from the NCBI SRA; isolates were selected from a recent publication that compiled isolates from several previous *C. difficile* genome collections along with their ST and sampling date [32]. The combined collection of existing and new sequences was then pared down to facilitate running Bayesian phylogenomic analyses. First, in an effort to maximize genetic diversity, one randomly selected isolate from each pair of isolates within 2 SNVs of one another was removed. Second, isolates from overrepresented geographic locations were randomly downsampled until the total number of isolates was less than 425. The final list of isolates that were included in these analyses can be found in Supplementary Table 1.

We assessed the suitability of the data for timed phylogenomic analyses by examining temporal signal—or the relationship between genomic differences and sampling date—using two methods. First, we examined a regression of sampling time vs. root-to-tip genetic distance using Tempest and BactDating [33,34]. We then formally evaluated temporal signal using date randomization tests, randomly permuting the sampling dates 10 times and comparing the 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 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

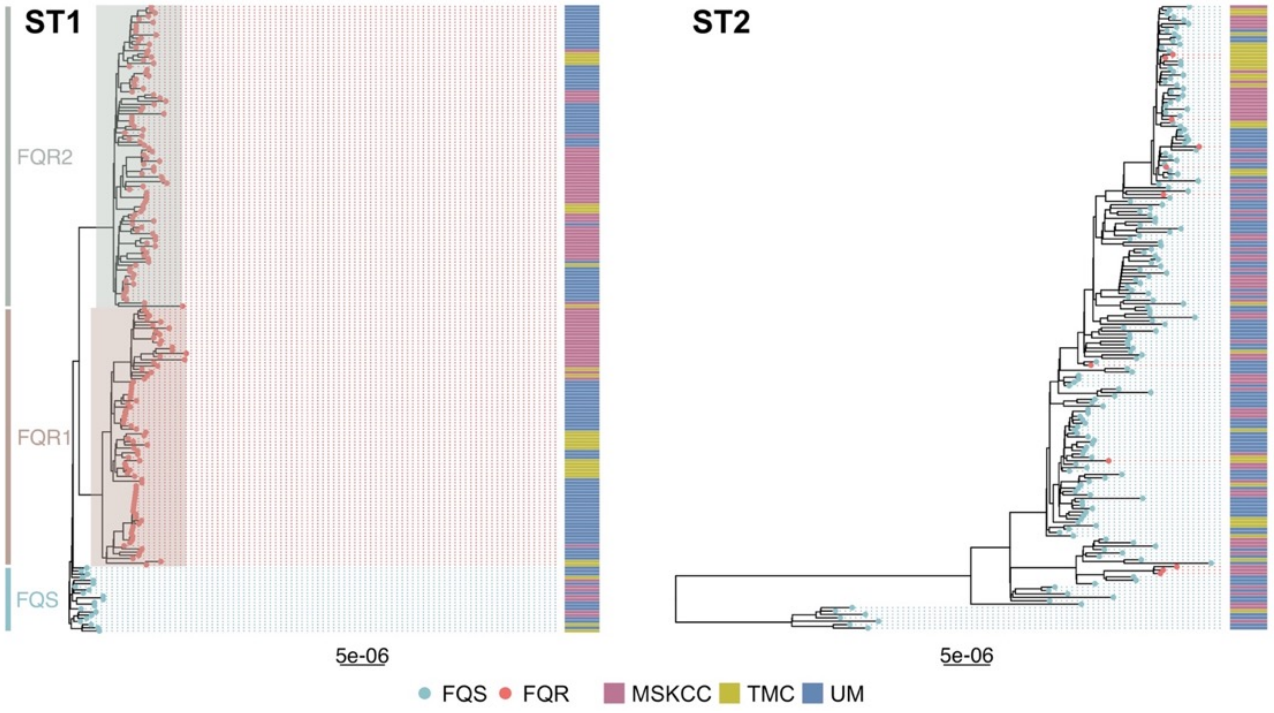
197 randomized datasets (CR1), and the more strict being met if the 95% credible interval of the
198 estimated evolutionary rate did not overlap any of the 95% credible intervals of the date
199 randomized datasets (CR2) [35]. We proceeded with evolutionary rate estimates so long as the
200 data met 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
208 warranted. To assess whether the data violated a strict clock assumption, we evaluated
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 5.0×10^{-7}
213 and standard deviation of 8×10^{-7} based on previous evolutionary rate estimates (while
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

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].

RESULTS

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

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 (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 nucleotide changes per quality- and recombination-filtered site.



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

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

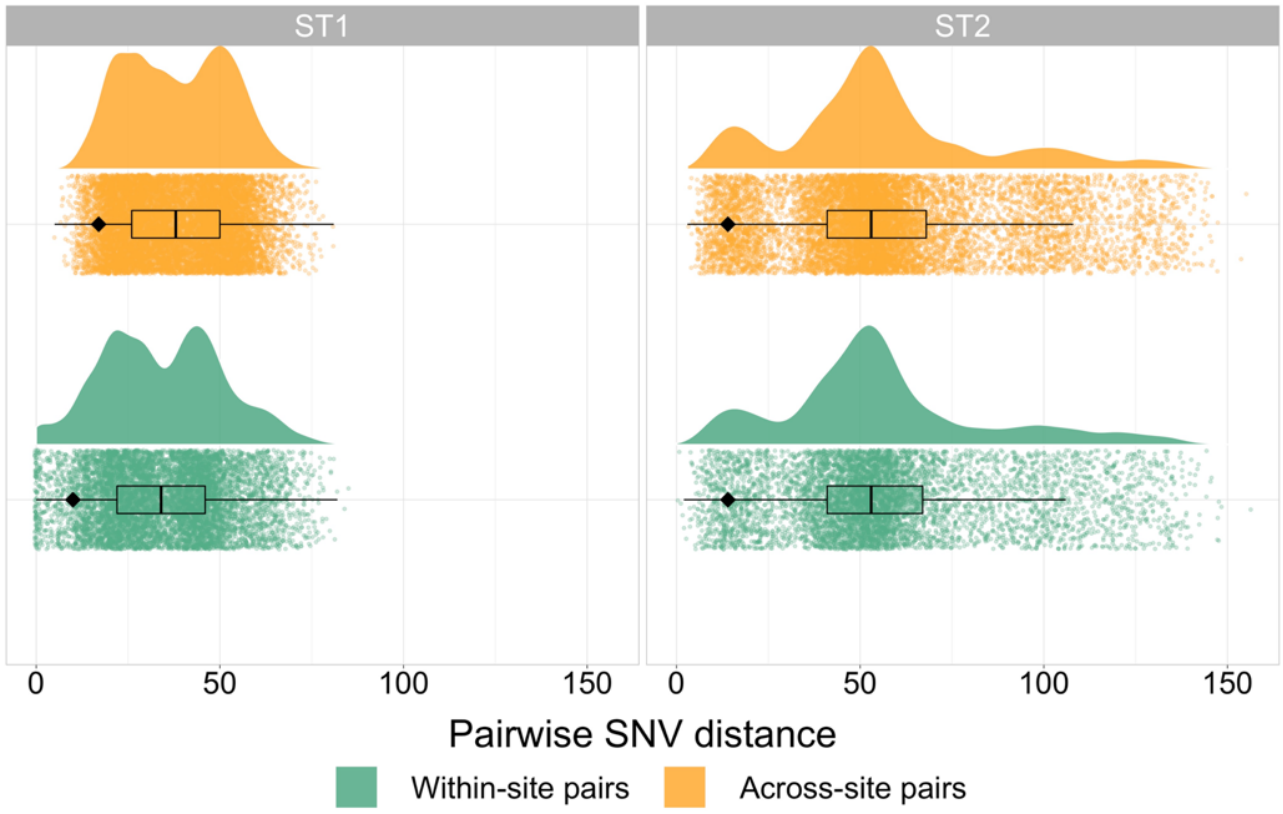
ST1 isolates display more evidence of recent transmission than ST2, while ST2 isolates are more likely to share intermediate genetic linkages across disparate geographic sites

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

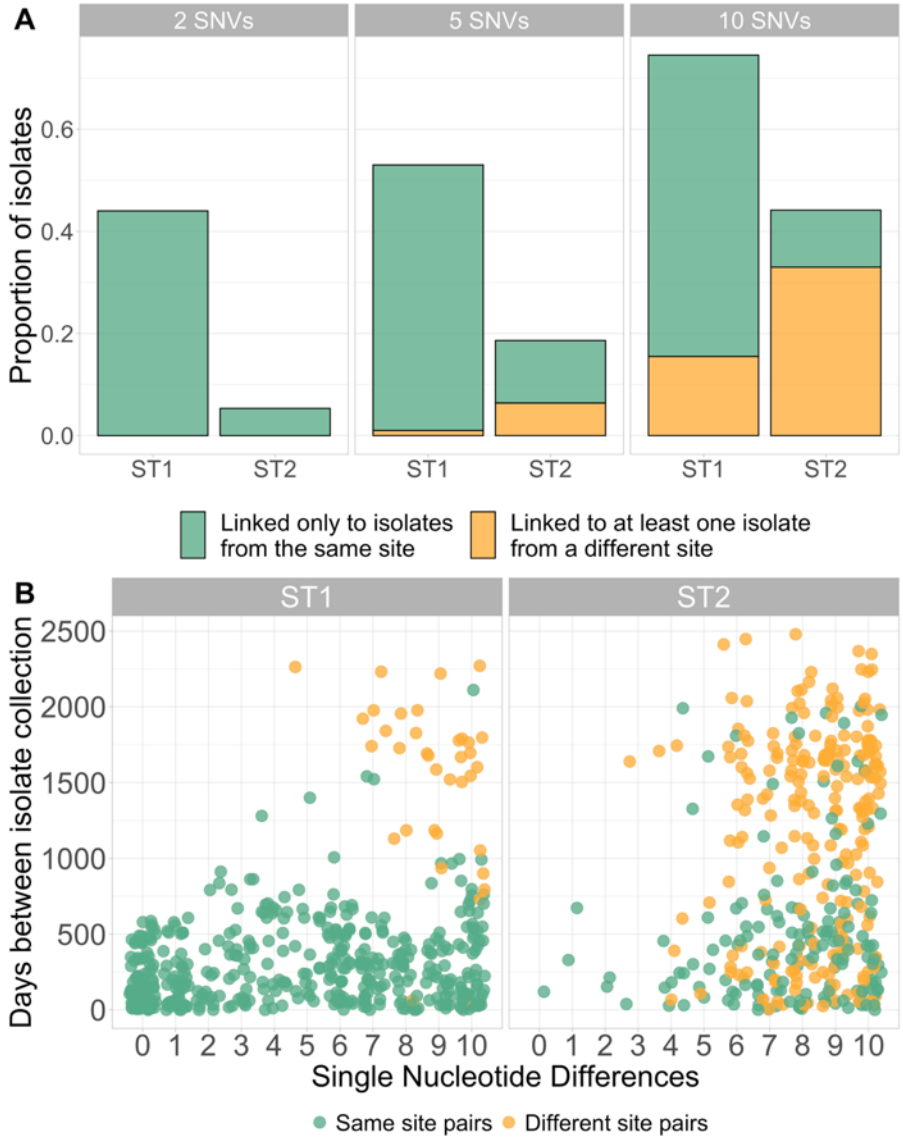
sample healthcare-associated transmission than community-associated transmission. When examining the SNV distance distributions between and within collection sites, among ST1 isolates, we observed more closely related pairs of isolates from the same geographic collection site (reflected by a heavier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

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

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 ST1 and ST2. The black diamond indicates fifth percentile SNV distances for each category.



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



301

Timed phylogenomic analyses demonstrate evidence of evolutionary rate heterogeneity within and between ST1 and ST2 lineages

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

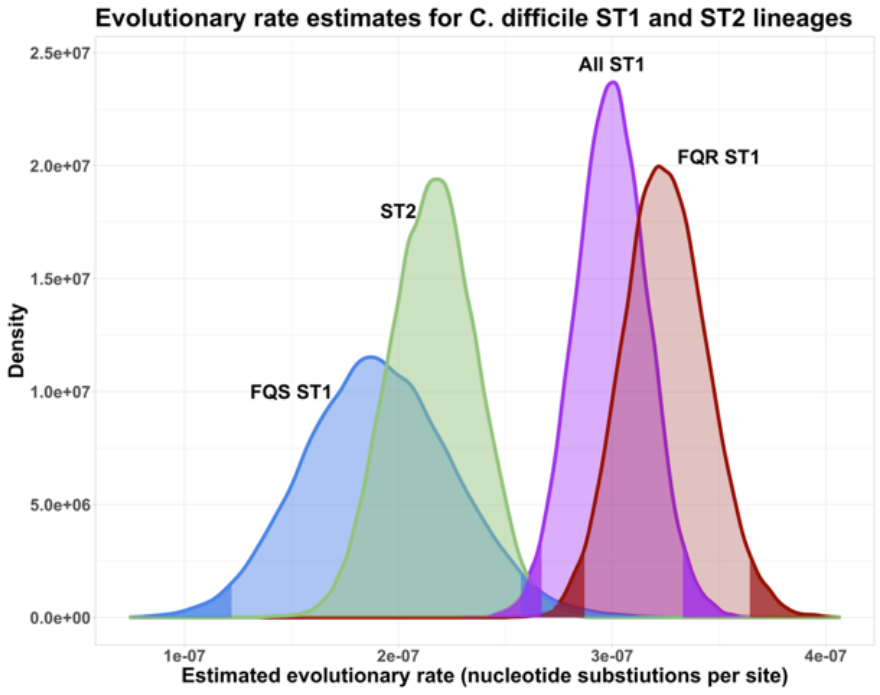
Temporal signal analyses, while initiated as a necessary precursor to timed phylogenomic analyses in BEAST, revealed interesting differences between the clock-like nature of ST1 and ST2 isolates. While root-to-tip regression analyses suggested similarly weak but sufficient temporal signal to proceed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

randomization tests demonstrated more evidence of temporal signal among ST1 isolates, which passed both the more relaxed CR1 and more stringent CR2 criteria for temporal analyses, compared to ST2 isolates, which passed CR1 but not CR2 (Supplementary Figure 4). The root-to-tip regression also highlighted different temporal patterns among FQS-ST1 isolates compared to FQR-ST1 isolates, which was observed again in date randomization tests on FQS-ST1 and FQR-ST1 isolates separately; the FQR-ST1 isolates appeared to drive the temporal signal in the data, and when considered alone, FQS-isolates were more like ST2 isolates, passing the more relaxed CR1 temporal signal criteria but not the more stringent CR2. This observation was consistent with our pairwise SNV distance findings of distinct patterns among FQS ST1 isolates, and motivated conducting further analyses both with all ST1 isolates together as well as with FQR ST1 isolates (n = 359) and FQS ST1 isolates (n = 59) considered separately.

All datasets demonstrated evidence of evolutionary rate heterogeneity throughout the phylogeny, resulting in the application of uncorrelated relaxed lognormal molecular clock models along with a constant demographic priors (see Supplementary Figures 5-6 and Supplementary Results for details). Overall, when considering all ST1 isolates together compared to all ST2 isolates, evolutionary rate estimates were slightly higher for ST1 compared to ST2, although the 95% credible intervals overlapped. However, ST1's faster evolutionary rate 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-overlapping 95% credible intervals) while FQS ST1 isolates had similar evolutionary rate estimates to ST2 isolates (Figure 4). These evolutionary rates translate to approximately 1.36

(95% credible interval 1.20-1.52) nucleotide changes per year for FQR ST1, 0.80 (95% credible interval 0.51-1.08) nucleotide changes per year for FQS-ST1, and 0.89 (95% credible interval 0.74-1.05) nucleotide changes per year for ST2. These results are consistent with the hypothesis that a slightly slower average evolutionary rate among ST2 and FQS ST1 isolates compared to FQR ST1 isolates might contribute to our observed discordance between genomic and epidemiologic linkages among those isolates.

Figure 4: Posterior probability density of the evolutionary rates estimates for *C. difficile* ST1 and ST2 lineages, with ST1 isolates considered together as well as separated out into FQR-ST1 and FQS-ST1 isolates. Dark shaded areas of the density curves indicate the lower 2.5% and upper 97.5% of the distributions; light shaded areas indicate 95% credible intervals. Evolutionary rates are considered significantly different from one another when the 95% credible intervals of their posterior probability densities do not overlap.



DISCUSSION

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

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

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 been reported as one of the most common strains In Europe [41], the US [13,42,42], and Australia [43] during the last decade. ST2 is commonly characterized in the literature as an endemic strain in the U.S. that has not been associated with hospital outbreaks [44]. However, it is also frequently classified as healthcare associated: the most recent data from the Centers 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]. Despite this, evidence of transmission of RT014/ST2 within the hospital is sparse, as demonstrated by this study and others [6,13]. One explanation for this discordance between genomic evidence of recent transmission and healthcare-associated characterization via surveillance definitions is that ST2 is frequently acquired in the community, imported into the hospital, and subsequently progresses to infection after hospitalization. If this is the case, antimicrobial stewardship interventions may be particularly effective for preventing infections due to this common strain [7]. Environmental studies that have reported recovery of RT014 isolates in agriculture [45,46], wastewater [47], and parks and homes [39] are also consistent 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

pathogen whole genome sequencing, genomic evidence of healthcare transmission could be used as an alternative and more accurate metric than infection onset for measuring within-hospital transmission of *C. difficile*.

We also observed a notable difference in concordance between genomic linkages (isolate related within small SNV distance thresholds) and epidemiologic linkages (isolates collected from the same site within temporally proximate time periods) among ST1 and ST2 isolates. Specifically, ST2 isolates were more likely to have close genomic neighbors across disparate geographic sites and long time periods. Consistent with this, a pan-European surveillance study reported that the average most closely related strain to any given RT014 isolate was collected from hundreds of miles away [13]. The mechanisms behind this finding are not clear, but are 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 isolates of the same species. In particular, it emphasizes the importance of considering background genomic diversity and incorporating geographically and temporally diverse strains when interpreting genomic linkages. Without this context, one might mistakenly attribute a linkage to transmission when it in fact reflects broader genomic diversity patterns in a particular lineage. The importance of genomic context has been noted since the early days of bacterial genomic epidemiology [48], but in most cases, sequencing is still not widespread enough to provide such context. As we continue to consider a future with routine genomic surveillance in hospital settings to identify outbreaks [49], it is crucial that assessment of

426 genomic context remain part of the evidence required for inferring transmission from genomic
427 data.
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
434 and FQS ST1 lineages may have spent, on average, more time in spore-form than the epidemic
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
437 pressures and a higher density of susceptible hosts in healthcare settings could facilitate more
438 time in the vegetative state, whereas strains that circulate primarily in the community may be
439 more likely to stay dormant for longer periods of time. Results from our Bayesian timed
440 phylogenomic analyses were consistent with this framework in two ways: 1) high evolutionary
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
443 with a slow estimated evolutionary rate and 2) less evidence of temporal signal and slightly
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

447 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
450 Our findings should be interpreted in the context of multiple limitations. First, the retrospective
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
459 emphasizes the value of integrating genomic data with even limited epidemiologic metadata.
460 Finally, the evolutionary rate estimates presented here are subject to uncertainty, particularly
461 given the observed instances of violated model assumptions and relatively limited temporal
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
464 notable strengths, including the collection of isolates from three distinct geographic sites in the
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.
468

Conclusions

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

ACKNOWLEDGEMENTS

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

FUNDING

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

CONFLICTS OF INTEREST

The authors declare that there are no conflicts of interest.

REFERENCES

1. 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.

2. 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.

3. Martin JSH, Monaghan TM, Wilcox MH. Clostridium difficile infection: Epidemiology, diagnosis and understanding transmission. Nat Rev Gastroenterol Hepatol. **2016**; 13(4):206–216.

4. Eyre DW, Cule ML, Wilson DJ, et al. Diverse sources of C. difficile infection identified on whole-genome sequencing. N Engl J Med. **2013**; 369(13):1195–205.

5. 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.

6. 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.

7. 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.

8. 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.

9. 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.

10. McDonald LC, Owens RC, Johnson S. An Epidemic, Toxin Gene–Variant Strain of Clostridium difficile. N Engl J Med. **2005**; 353(23):2433-2441.

11. 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.

12. 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.

13. Guh AY, Mu Y, Winston LG, et al. Trends in U.S. Burden of *Clostridioides difficile* Infection and Outcomes. N Engl J Med. **2020**; 382(14):1320–1330.

14. 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.

15. Gonzales-Luna AJ, Carlson TJ, Dotson KM, et al. PCR ribotypes of *Clostridioides difficile* across Texas from 2011 to 2018 including emergence of ribotype 255. Emerg Microbes Infect. **2020**; 9(1):341–347.

16. 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*.

17. Martinson JNV, Broadaway S, Lohman E, et al. Evaluation of Portability and Cost of a Fluorescent PCR Ribotyping Protocol for Clostridium difficile Epidemiology. Onderdonk AB, editor. J Clin Microbiol. **2015**; 53(4):1192–1197.

18. Griffiths D, Fawley W, Kachrimanidou M, et al. Multilocus Sequence Typing of *Clostridium difficile*. J Clin Microbiol. **2010**; 48(3):770–778.

19. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. Bioinformatics. **2014**; 30(15):2114–2120.

20. He M, Sebaihia M, Lawley TD, et al. Evolutionary dynamics of Clostridium difficile over short and long time scales. Proc Natl Acad Sci. **2010**; 107(16):7527–7532.

21. Yin C, Chen DS, Zhuge J, et al. Complete Genome Sequences of Four Toxigenic *Clostridium difficile* Clinical Isolates from Patients of the Lower Hudson Valley, New York, USA. Genome Announc. **2018**; 6(4):e01537-17.

22. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics. **2009**; 25(14):1754–1760.

23. Li H, Handsaker B, Wysoker A, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. **2009**; 25(16):2078–2079.

24. Croucher NJ, Page AJ, Connor TR, et al. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Res. **2015**; 43(3):e15.

25. Hunt M, Mather AE, Sánchez-Busó L, et al. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. Microb Genomics. **2017**; 3(10):e000131.

557 26. Nguyen LT, Schmidt HA, Von Haeseler A, Minh BQ. IQ-TREE: A fast and effective stochastic
558 algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol. **2015**; 32(1):268–
559 274.

560 27. Sebaihia M, Wren BW, Mullany P, et al. The multidrug-resistant human pathogen
561 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.

565 29. He M, Miyajima F, Roberts P, et al. Emergence and global spread of epidemic healthcare-
566 associated Clostridium difficile. Nat Genet. **2013**; 45(1):109–113.

567 30. Popovich KJ, Snitkin ES, Hota B, et al. Genomic and Epidemiological Evidence for
568 Community Origins of Hospital-Onset Methicillin-Resistant Staphylococcus aureus
569 Bloodstream Infections. J Infect Dis. **2017**; 215:1640–1647.

570 31. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, Rambaut A. Bayesian
571 phylogenetic and phylodynamic data integration using BEAST 1.10. Virus Evol. **2018**;
572 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.

576 33. Rambaut A, Lam TT, Max Carvalho L, Pybus OG. Exploring the temporal structure of
577 heterochronous sequences using TempEst (formerly Path-O-Gen). Virus Evol. **2016**;
578 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.

584 36. Didelot X, Eyre DW, Cule M, et al. Microevolutionary analysis of Clostridium difficile
585 genomes to investigate transmission. Genome Biol. **2012**; 13(12):R118.

586 37. Minin VN, Bloomquist EW, Suchard MA. Smooth Skyride through a Rough Skyline:
587 Bayesian Coalescent-Based Inference of Population Dynamics. Mol Biol Evol. **2008**;
588 25(7):1459–1471.

589 38. Rambaut A, Suchard M, Xie D, Drummond A. Tracer v1.6. **2014**; . Available from:
590 http://beast.bio.ed.ac.uk/Tracer

591 39. Alam MJ, Walk ST, Endres BT, et al. Community Environmental Contamination of Toxigenic
592 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.

600 42. Tenover FC, Tickler IA, Persing DH. Antimicrobial-Resistant Strains of Clostridium difficile
601 from North America. Antimicrob Agents Chemother. **2012**; 56(6):2929–2932.

602 43. Foster NF, Collins DA, Ditchburn SL, et al. Epidemiology of Clostridium difficile infection in
603 two tertiary-care hospitals in Perth, Western Australia: a cross-sectional study. New
604 Microbes New Infect. **2014**; 2(3):64–71.

605 44. Aitken SL, Alam MJ, Khaleduzzuman M, et al. In the Endemic Setting, *Clostridium difficile*
606 Ribotype 027 Is Virulent But Not Hypervirulent. Infect Control Hosp Epidemiol. **2015**;
607 36(11):1318–1323.

608 45. Knight DR, Squire MM, Collins DA, Riley TV. Genome Analysis of Clostridium difficile PCR
609 Ribotype 014 Lineage in Australian Pigs and Humans Reveals a Diverse Genetic Repertoire
610 and Signatures of Long-Range Interspecies Transmission. Front Microbiol. **2017**; 7:2138

611 46. Janezic S, Zidaric V, Pardon B, et al. International Clostridium difficile animal strain
612 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.

618 49. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection
619 by leading with genomic surveillance of nosocomial pathogens. Microbiology. **2018**;
620 164(10):1213–1219.

621 50. Weller C, Wu M. A generation-time effect on the rate of molecular evolution in bacteria.
622 Evolution. **2015**; 69(3):643–652.

623