MICROBIAL GENOMICS

RESEARCH ARTICLE

Miles-Jay *et al.*, *Microbial Genomics* 2021;7:000590

DOI 10.1099/mgen.0.000590

MICROBIOLOGY SOCIETY

OPEN DATA OPEN ACCESS

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

Arianna Miles-Jay[1], Vincent B. Young[1], Eric G. Pamer[2,3], Tor C. Savidge[4], Mini Kamboj[2,5], Kevin W. Garey[6] and Evan S. Snitkin[1,*]

Abstract

Clostridioides difficile is the leading cause of healthcare-associated infectious diarrhoea. However, it is increasingly appreciated that healthcare-associated infections derive from both community and healthcare environments, 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 USA: sequence type (ST) 1 (associated with ribotype 027) and ST2 (associated with ribotype 014/020). We performed whole-genome sequencing and phylogenetic analyses on 382 ST1 and ST2 *C. difficile* isolates recovered from stool specimens collected during standard clinical care at 3 geographically distinct US medical centres between 2010 and 2017. 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 two SNVs of another isolate from the same geographical collection site compared to 5.5% of ST2 isolates (*P*-value=<0.001). Conversely, ST2 isolates were more likely to have close genetic neighbours across disparate geographical 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 geographical concordance among closely related isolates. Together, these findings suggest that 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-specific criteria for interpreting genomic linkages and emphasizes the importance of considering differences in the epidemiology of circulating strains when devising interventions to reduce the burden of *C. difficile* infections.

DATA SUMMARY

All whole-genome sequence data were uploaded to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProject accessions PRJNA595724, PRJNA561087 and PRJNA594943. Metadata that comply with patient privacy rules are included in the Supplementary Material.

Received 10 December 2020; Accepted 19 April 2021; Published 28 June 2021

Author affiliations: [1]University of Michigan Medical School, Ann Arbor, MI, USA; [2]Memorial Sloan Kettering Cancer Center, New York, NY, USA; [3]The University of Chicago, Chicago, IL, USA; [4]Baylor College of Medicine, Houston, TX, USA; [5]Weill Cornell Medical College, New York, NY, USA; [6]University of Houston College of Pharmacy, Houston, TX, USA.

***Correspondence:** Evan S. Snitkin, esnitkin@umich.edu

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

Abbreviations: FQR, fluoroquinolone resistant; FQS, fluoroquinolone sensitive; GMRF, Gaussian Markov random field; HPD, highest posterior density interval; MLST, multilocus sequence typing; MSKCC, Memorial Sloan Kettering Cancer Center; NCBI, National Center for Biotechnology Information; SNV, single nucleotide variant; SRA, sequence read archive; ST, sequence type; TMC, Texas Medical Center Hospitals; UM, Michigan Medicine; WGS, whole genome sequencing.

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

Data statement: All supporting data, code and protocols have been provided within the article or through supplementary data files. One supplementary table and six supplementary figures are available with the online version of this article.

INTRODUCTION

Clostridioides difficile is a Gram-positive spore-forming anaerobic bacterium that is a dominant cause of infectious diarrhoea, colitis and colitis-associated death in the USA [1, 2]. While *C. difficile* infection (CDI) is classically considered to be nosocomial [3], recent molecular epidemiological research suggests that fewer 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 time since hospitalization [9–12]. Another common *C. difficile* lineage in the USA, 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 meaningful metric for assessing 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 epidemiological patterns have not yet been assessed using genomic data gathered from USA-based *C. difficile* isolates. Here, we applied WGS to isolates collected from three geographically distinct US medical centres to assess differential genomic evidence of healthcare vs community spread between two of the most common *C. difficile* strains: ST1 and ST2.

Impact Statement

Clostridioides difficile is a leading cause of healthcare-associated infections, and new strategies for preventing *C. difficile* infections are urgently needed. However, there are many different strains of *C. difficile,* and existing evidence suggests that some strains may spread more frequently within healthcare settings while some may spread more frequently in the community. Whole-genome sequencing of *C. difficile* isolates from multiple geographical locations with diverse patient populations can shed light on which strains spread more readily within hospitals, but data from the USA are lacking. This study leverages a collection of whole-genome sequences from three geographically distant sites in the USA and demonstrates genomic evidence for differential reliance on healthcare vs community spread between two of the most common strains causing *C. difficile* infection: ST1 and ST2. Furthermore, we highlight potential pitfalls in analysing *C. difficile* genomic data without appropriate geographical context and an understanding of strain-specific genetic diversity. These findings support the potential application of different prevention strategies depending on local molecular epidemiology of *C. difficile* infections, and raise interesting questions about the epidemiological and biological underpinnings of why some types of *C. difficile* seem to spread more in healthcare while others may spread more in the community.

METHODS

Data collection

C. difficile sequences were derived from clinical stool specimens collected as part of existing molecular surveillance programmes that took place at three US medical centres: 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 specimens and DNA was extracted from a single colony using the Qiagen MagAttract Microbial DNA kit (Qiagen, Inc., Germantown, MD, USA) at UM; either the QIAamp DNA mini kit (Qiagen, Inc., Germantown, MD, USA) or the AnaPrep automated DNA extractor (BioChain Institute, Inc., Newark, CA, USA) at TMC; and the QIAamp DNA mini kit at MSKCC (Qiagen, Inc., Germantown, MD, USA) as previous described [14, 16, 17]. Isolates underwent fluorescent PCR ribotyping at UM and TMC [18] and MLST at MSKCC [19]. DNA from a sample of isolates that were typed as RT027 at UM and TMC or ST1 at MSKCC (*n*=201) and RT014/020 at UM and TMC or ST2 at MSKCC (*n*=263) was sent to UM for WGS. The Institutional Review Boards at each of the study sites approved the study protocols.

WGS and bioinformatic methods

The Nextera XT library preparation kit (Illumina, San Diego, CA, USA) 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 >100×. Sequence data are available from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under BioProjects PRJNA595724, PRJNA561087 and PRJNA594943. First, *in silico* multilocus sequence typing (MLST) was performed on the raw sequencing reads using ARIBA; only isolates that were identified as ST1 and ST2 were included in further analyses (82 sequences were excluded) [19–21]. The bioinformatics methods applied to the *C. difficile* sequences to identify single-nucleotide variants (SNVs) and build phylogenetic trees were executed as previously described [22]. Briefly, raw sequencing reads were trimmed using Trimmomatic to remove low-quality bases and adapter sequences [23]. 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 [24–26]. PCR duplicates were discarded and variants were called using SAMtools mpileup and bcftools [27]. Gubbins was used to remove variant sites located in putative recombinant regions [28]. 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) [29, 30]. Fluroquinolone resistance was inferred based on the presence of previously identified fluroquinolone resistance-associated *gryA* and *gyrB* alleles [31]. ST1 isolates were further classified into previously identified FQS, FQR1 and FQR2 lineages by examining how newly sequenced isolates clustered with publicly available FQR1 and FQR2 isolates [32].

Evaluation of phylogenetic clustering

To compare the level of clustering by geographical collection site between newly sequenced ST1 and ST2 isolates, we overlaid geographical collection site onto the maximum-likelihood whole-genome phylogenies and applied a previously described approach for formal clustering assessment [33]. First, we tabulated the number of isolates in a 'pure' subtree of each phylogeny – defined as a subtree made up of two or more isolates collected from a single geographical 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 geographical labels and recalculating this number 1000 times.

Evaluation of evidence of recent transmission

Evidence of recent transmission was assessed using pairwise SNV distance matrices and two analytic approaches. First, we compared the lower tail (fifth percentile) of the distribution of pairwise SNV distances of pairs of isolates collected from the same collection site to that same metric among pairs of isolates collected from different collection sites by calculating a fifth percentile SNV distance ratio (fifth percentile SNV distance within sites/fifth percentile SNV distance between sites). To assess whether this ratio indicated an enrichment of close linkages within collection sites greater than could be expected by chance, we randomly permuted collection sites and recalculated the ratio 10 000 times; an observed ratio below the 2.5% percentile of the distribution of expected ratios was applied to support significant enrichment of close genetic linkages within study sites. Second, we classified genomic linkages using an SNV distance threshold of two SNVs and compared the proportion of genomically linked isolates (defined as being linked to at least one other isolate) among ST1 isolates compared to those among ST2 isolates using chi-squared tests. An SNV threshold of two SNVs is commonly used to identify pairs of *C. difficile* isolates that are likely related via direct transmission/acquisition from a common source; this threshold is based on evolutionary rates estimated from within-host evolution [4]. We then assessed the sensitivity of these results to larger thresholds of 5–10 SNVs. We also compared the proportion of isolates genomically linked to at least one isolate collected from a different geographical collection site between ST1 and ST2 using chi-squared tests. All analyses were completed in R v4.0.2 (R Core Team, 2020).

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 [34]. 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 [35]. 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 two SNVs of one another was removed. Second, isolates from overrepresented geographical locations were randomly downsampled until the total number of isolates was <425. The final list of isolates that were included in these analyses can be found in Table S1 (available in the online version of this article).

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 [36, 37]. 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 the stricter criteria for temporal signal assessment

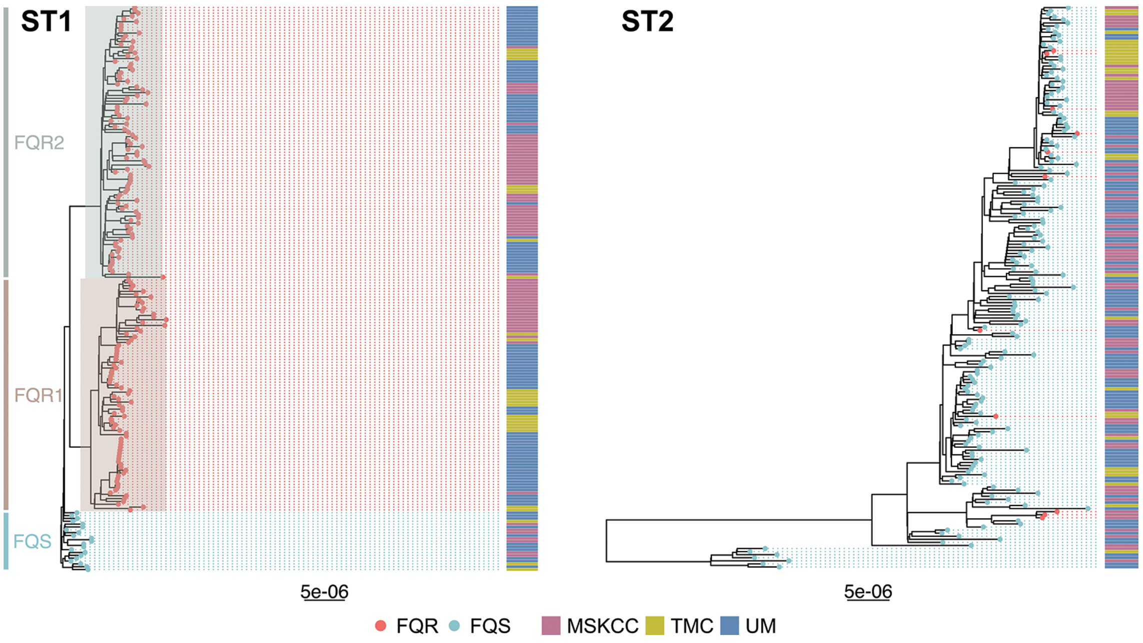

Fig. 1. Maximum-likelihood phylogenetic trees of newly sequenced *C. difficile* isolates that are ST1 and ST2. Tips are coloured by fluroquinolone-resistant (FQR) vs fluroquinolone-sensitive (FQS) as determined by the presence of previously identified fluroquinolone-resistance-associated *gryA* and *gyrB*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

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 randomized datasets (CR1), and the stricter 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) [38]. We proceeded with evolutionary rate estimates so long as the data met CR1.

To select BEAST model assumptions for both the date randomization tests and the final evolutionary rate estimates, we started with a general time-reversible nucleotide substitution model with gamma distributed rate heterogeneity and the simplest clock and demographic model assumptions: a strict molecular clock and constant demographic prior. We then systematically examined the extent to which the data violated the strict clock and constant demographic model prior assumptions and thus the extent to which more complex models were warranted. To assess whether the data violated a strict clock assumption, we evaluated whether the coefficient of variation parameter in the models with an uncorrelated relaxed lognormal clock had a 95% highest posterior density interval (HPD) that overlapped 0; if not, we used this as evidence of the assumptions of a strict clock being violated and applied an uncorrelated relaxed lognormal clock with a lognormal prior distribution with a mean of 5.0×10^{-7} and standard deviation of 8×10^{-7} based on previous evolutionary rate estimates (while still allowing for significant deviation) [32, 39]. To assess the extent to which the data violated a constant demographic model, we ran models with exponential growth demographic model prior, and evaluated whether the 95% credible interval of the exponential growth rate parameter overlapped 0. If the exponential growth rate parameter was substantially different 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 [40]. For each model, a Markov chain Monte Carlo was run for 200 million generations and sampled every 10000 iterations; a Tempest-rooted starting tree was included in all runs to accelerate convergence [36]. All ESS values were checked for being above 200 using Tracer after removing the first 10% of steps as burn-in [41].

RESULTS

Three hundred and eighty-two new whole-genome sequences were generated from the 3 US study sites located in Michigan, Texas and New York; 199 ST1 and 183 ST2 (Fig. S1). 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 (Fig. 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) (Fig. 2). The ST1 phylogeny

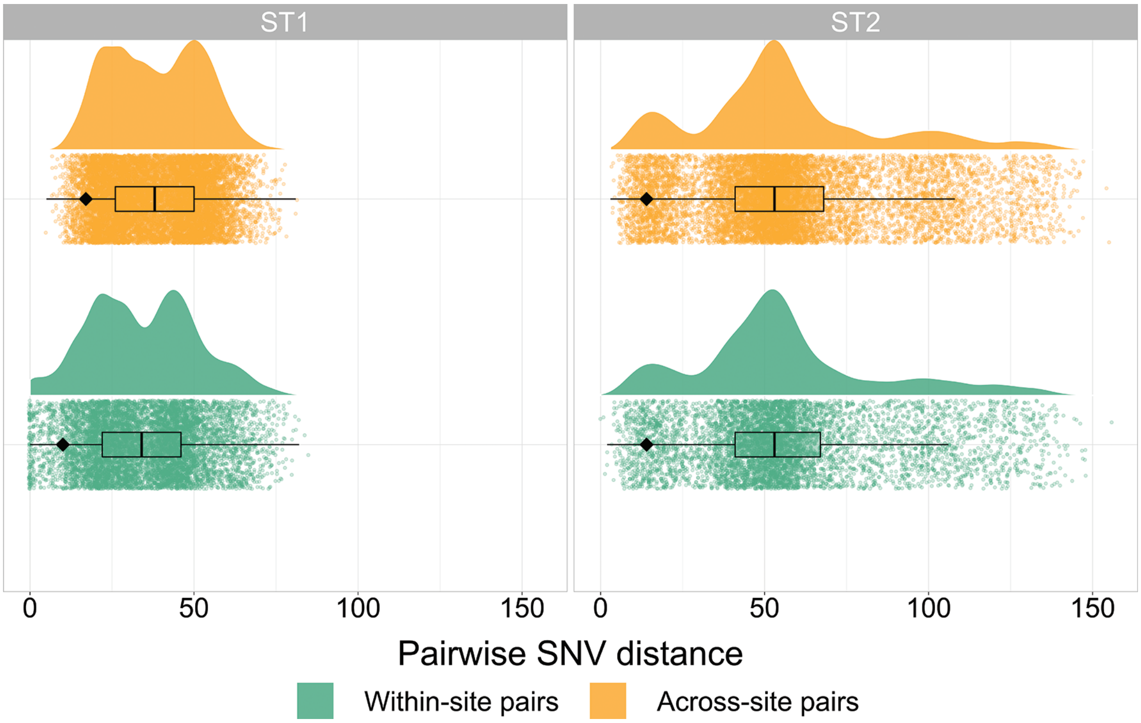

Fig. 2. Pairwise single-nucleotide variant (SNV) distribution between pairs of isolates from the same collection site vs pairs of isolates from geographically distinct collection sites for both ST1 and ST2. The black diamond indicates fifth percentile SNV distances for each category.

displayed shorter terminal branches than the ST2 phylogeny, which is also consistent with more genetic similarity among sampled ST1 isolates compared to ST2 isolates.

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 geographical structure by overlaying the geographical site each isolate was collected from onto strain-specific whole-genome phylogenies. Visual examination of these phylogenies revealed a striking difference in geographical clustering, with ST1 displaying larger clusters and ST2 displaying more numerous, smaller clusters and more geographical mixing (Fig. 1). The exception to this observation was the FQS ST1 clade, which appeared to be more geographically mixed than the FQR ST1 clades. While statistical assessments demonstrated that both ST1 and ST2 displayed more evidence of geographical clustering than would be expected to occur by chance (empirical *P*-values both <0.001), clustering was more non-random for ST1 than ST2 (Fig. S2). This enhanced geographical clustering among ST1 isolates could reflect the fact 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 geographical sites

To further investigate whether plausible healthcare-associated transmission among ST1 isolates was driving the geographical 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 geographical collection site (reflected by a heavier lower tail of the distribution) compared to pairs of isolates collected from different geographical collection sites (fifth percentile SNV distance within sites/fifth percentile SNV distance between sites=0.59, expected ratio 95% interval 0.93–1.00, Fig. 2). However, we did not observe this same pattern among ST2 isolates (fifth percentile SNV distance within sites/fifth percentile SNV distance between sites=1.00, expected ratio 95% interval 0.93–1.00, Fig. 2). Application of SNV distance thresholds demonstrated that 88 (44%) ST1 isolates were within 2 SNVs of another isolate from the

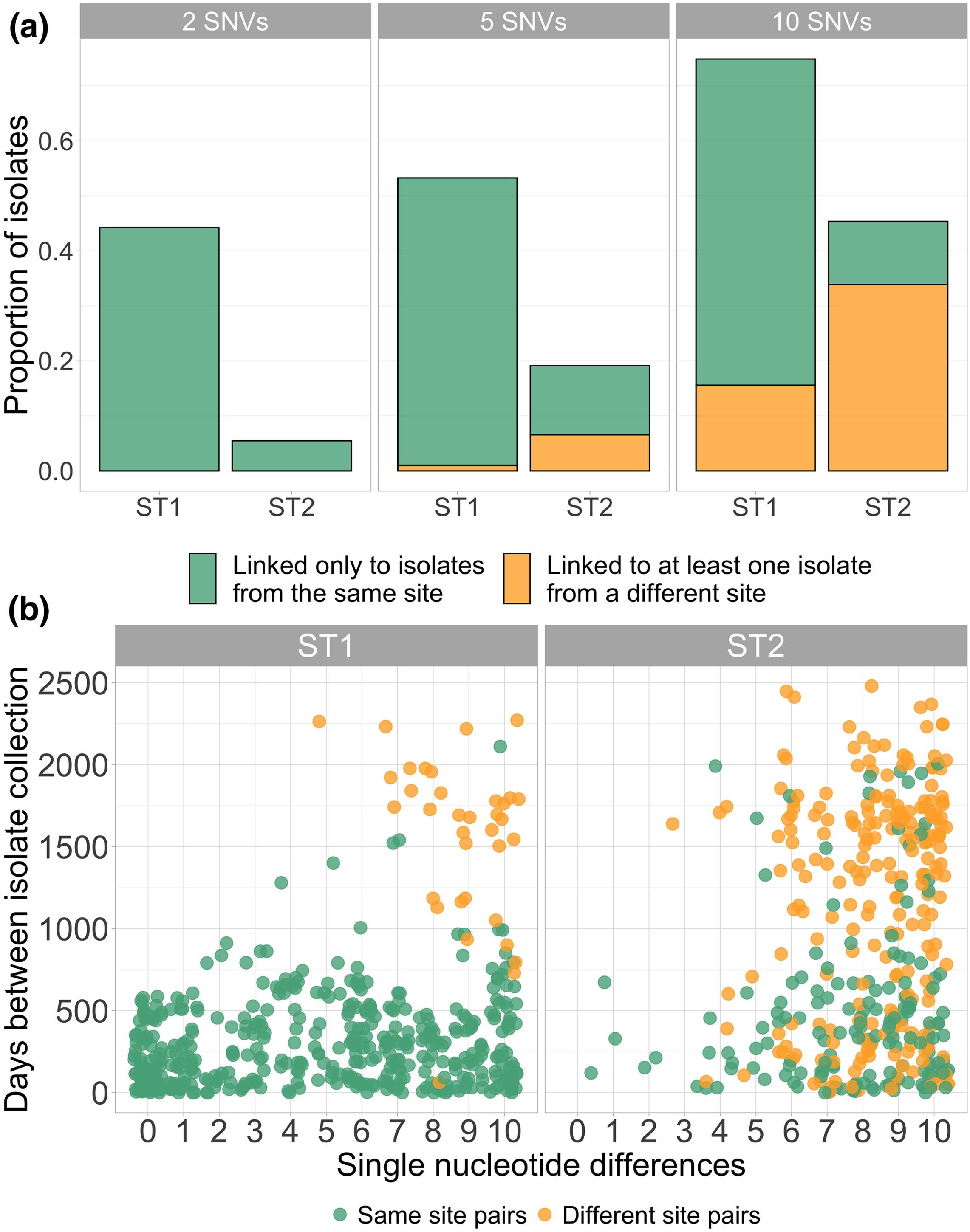

Fig. 3. (a) Bar plot showing the proportion of ST1 and ST2 isolates that are genomically linked to another isolate, either from the same collection site only (green) or from at least one different collection site (orange), at varying SNV thresholds. (b) Scatter plot of days between collection and pairwise SNV distance up to 10 SNVs, where each dot represents 1 pair of isolates. Points are coloured by whether they are collected from the same geographical collection site (green) or different geographical collection sites (orange). Points are jittered to improve clarity.

same geographical 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, Fig. 3a). Conversely, at the 5 and 10 SNV thresholds, linked ST2 isolates were more likely to be linked to an isolate from a different geographical collection site compared to linked ST1 isolates (all *P* <0.001, Fig. 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 2479 days (Fig. 3b). Among geographically discordant ST1 isolate 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 geographical sites.

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 geographical 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 geographical distances and/or (2) a slower average evolutionary rate among ST2 isolates resulting in fewer 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 BEAST Bayesian phylogenetic software [34]. Four hundred and eighteen ST1 and 418 ST2 isolates were 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 BEAST software (Table S1). 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, Fig. S3), 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 the more stringent CR2 criteria for temporal analyses, compared to ST2 isolates, which passed CR1 but not CR2 (Fig. S4). 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 Figs S5 and S6 and File S1 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 those of ST2 isolates (with non-overlapping 95% credible intervals), while FQS ST1 isolates had similar evolutionary rate estimates to ST2 isolates (Fig. 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 epidemiological linkages among those isolates.

DISCUSSION

In this study, we investigated the genomic epidemiology of two dominant *C. difficile* lineages, ST1 and ST2, across three geographically distinct US medical centres. We observed more genomic evidence of geographical clustering and recent transmission among ST1 isolates compared to ST2 isolates, while also finding more linkages among ST2 isolates from disparate geographical 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 on a combination of newly sequenced USA-based isolates and publicly available global isolates.

Previous studies have reported both more evidence of broad geographical clustering [8] and more evidence of recent

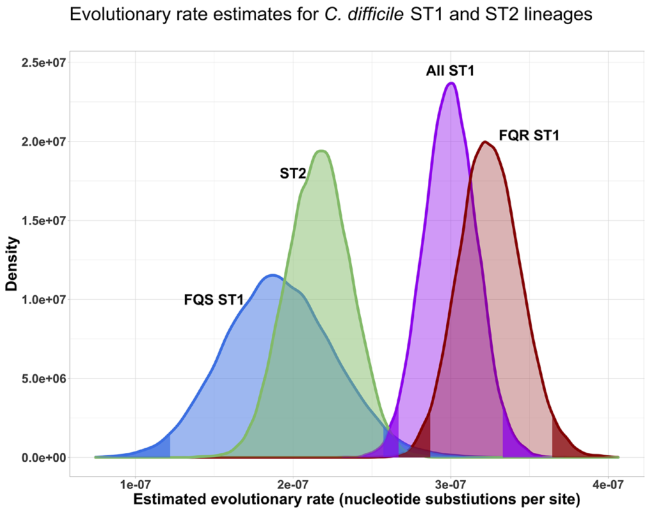

Fig. 4. Posterior probability density of the evolutionary rates estimates for *C. difficile* ST1 and ST2 lineages, with ST1 isolates considered together as well as separated out into FQR-ST1 and FQS-ST1 isolates. Dark shaded areas of the density curves indicate the lower 2.5% and upper 97.5% of the distributions; 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.

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 USA-based multisite data to support these findings. Our observations are consistent with ST1 being associated with hospital outbreaks [10, 11, 42], being the most predominant healthcare-associated *C. difficile* strain according to surveillance definitions based on time since hospitalization [13], and being more prevalent in hospital than community environmental sampling [43]. The factors contributing to increased spread of ST1 within healthcare are not well defined, but 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 [44].

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 [45], the USA [13, 46] and Australia [47] during the last decade. ST2 is commonly characterized in the literature as an endemic strain in the USA that has not been associated with hospital outbreaks [48]. However, it is also frequently classified as healthcare associated: the most recent data from the Centers for Disease Control and Prevention Emerging Infections Program reported that 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 detected after hospitalization. If this is the case, antimicrobial and diagnostic stewardship interventions would be particularly important in settings where ST2 is common [7]. Environmental studies that have reported recovery of RT014 isolates in agriculture [49, 50], wastewater [51], and parks and homes [43] are also consistent with community circulation of RT014. Overall, this finding highlights the imperfect nature of relying on time since hospitalization as a proxy for acquisition. With the advent of more widespread pathogen WGS, genomic evidence of healthcare transmission and/or identification of known hospital-associated strains could be used as more reliable and actionable metrics to monitor and prevent hospital-acquired CDI.

We also observed a notable difference in concordance between genomic linkages (isolate related within small SNV distance thresholds) and epidemiological 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 neighbours across disparate geographical 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 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 [52], 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 [53], it is crucial that assessment of genomic context remains part of the evidence required for inferring transmission from genomic data.

C. difficile's spore-forming lifestyle may contribute to some of the results reported here. It has been posited that spore formation likely drags down average estimate evolutionary rates of bacteria [54]. Extending from that, if isolates belonging to particular lineages spend more time in spore form than others, that lineage could be expected to have a lower average evolutionary rate, and thus fewer 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 and more recently emerged FQR ST1 lineages, resulting in more closely related isolates across 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 time in the vegetative state, whereas strains that circulate primarily in the community may be more likely to stay dormant for longer periods of time. The results from our Bayesian timed phylogenomic analyses were consistent with this framework in two ways: (1) high evolutionary rate heterogeneity in both ST1 and ST2 isolates may reflect the effects of spore formation, with 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 lower estimated evolutionary rates for FQS-ST1 isolates and ST2 isolates compared to FQR-ST1 isolates may reflect more time spent in spore form. While limited experimental work has demonstrated no difference in spore-forming characteristics between RT027 and RT014/020 strains *in vitro*, this does not preclude differences at a population level [55]. Whatever the biological and epidemiological underpinnings of the patterns we observed, this work highlights the challenges inherent to applying molecular clock-based methods to studying the epidemiology and evolution of a variably and relatively slowly evolving pathogen like *C. difficile*.

Our findings should be interpreted in the context of multiple limitations. First, the retrospective nature of the study resulted in some differences in sample collection between the three study sites: UM and TMC selected based on PCR ribotypes, which we then filtered down to only ST1 and ST2 via *in silico* MLST, while MSKCC originally selected isolates based on ST as MLST is routine at that centre. However, all comparisons were made between ST1 and ST2 isolates and these differences were consistent within the ST1 and ST2 isolates at each site, so we would not expect them to significantly alter the results reported here. Second, limited epidemiological metadata were available for analysis, only study site and collection date, and thus we were not able to assess detailed epidemiological exposures or differences based on patient characteristics. Despite this, the interesting patterns we observed between genomic linkages and geography emphasize the value of integrating genomic data with even limited epidemiological metadata. Finally, the evolutionary rate estimates presented here are subject to uncertainty, particularly given the observed instances of violated model assumptions and relatively limited temporal signal in the data. However, the overall trends remained stable with varying models, alleviating concerns that our findings are artefacts of model misspecification. This study also has several notable strengths, including the collection of isolates from three distinct geographical sites in the USA, the application of WGS for high-resolution typing and phylogenetic analyses, and the incorporation of global isolates for increased context and power in our timed phylogenomic analyses.

Conclusions

Examination of the genomic epidemiology of *C. difficile* ST1 and ST2 across three geographically distinct US medical centres revealed divergent epidemiological and evolutionary patterns between these two common strains. Specifically, we observed more evidence of geographical 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 strategies 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.

Funding information

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.), the Cancer Centre Support Grant (P30CA008748, M.K), U01AI124290 (T.C.S., K.W.G.), and U01AI124275 (E.G.P., M.K.).

Acknowledgements

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

Conflicts of interest

The authors declare that there are no conflicts of interest.

References

1. Magill SS, O'Leary E, Janelle SJ, Thompson DL, Dumyati G, *et al*. Changes in prevalence of health care–associated infections in U.S Hospitals. *N Engl J Med* 2018;379:1732–1744.
2. Lessa FC, Mu Y, Bamberg WM, Beldavs ZG, Dumyati GK, *et al*. Burden of Clostridium difficile infection in the United States. *N Engl J Med* 2015;372:825–834.
3. Martin JSH, Monaghan TM, Wilcox MH. *Clostridium difficile* infection: Epidemiology, diagnosis and understanding transmission. *Nat Rev Gastroenterol Hepatol* 2016;13:206–216.
4. Eyre DW, Cule ML, Wilson DJ, Griffiths D, Vaughan A, *et al*. Diverse sources of *C. difficile* infection identified on whole-genome sequencing. *N Engl J Med* 2013;369:1195–1205.
5. Walker AS, Eyre DW, Wyllie DH, Dingle KE, Harding RM, *et al*. Characterisation of *Clostridium difficile* hospital ward-based transmission using extensive epidemiological data and molecular typing. *PLoS Med* 2012;9.
6. Martin JSH, Eyre DW, Fawley WN, Griffiths D, Davies K. Patient and strain characteristics associated with *Clostridium difficile* transmission and adverse outcomes. *Clin Infect Dis* 2018;1:9.
7. Poirier D, Gervais P, Fuchs M, Roussy JF, Paquet-Bolduc B, *et al*. Predictors of *Clostridioides difficile* infection among asymptomatic, colonized patients: a retrospective cohort study. *Clin Infect Dis* 2019:ciz626.
8. Eyre DW, Davies KA, Davis G, Fawley WN, Dingle KE, *et al*. Two distinct patterns of *Clostridium difficile* diversity across

Europe indicating contrasting routes of spread. *Clin Infect Dis* 2018;67:1035–1044.

9. Loo VG, Oughton M, Bourgault AM, Kelly M, Dewar K, *et al*. A predominantly clonal multi-institutional outbreak of *Clostridium difficile*–associated diarrhea with high morbidity and mortality. *N Engl J Med* 2005;353:2442–2449.
10. McDonald LC, Owens RC, Johnson S. An epidemic, toxin gene–variant strain of *Clostridium difficile*. *N Engl J Med* 2005;353:2433–2441.
11. Warny M, Pepin J, Fang A, Killgore G, Thompson A, *et al*. Toxin production by an emerging strain of *Clostridium difficile* associated with outbreaks of severe disease in North America and Europe. *Lancet* 2005;366:1079–1084.
12. Kuijper EJ, Barbut F, Brazier JS, Kleinkauf N, Eckmanns T, *et al*. Update of *Clostridium difficile* infection due to PCR ribotype 027 in Europe, 2008. *Eurosurveillance* 2008;13.
13. Guh AY, Mu Y, Winston LG, Johnston H, Olson D, *et al*. Trends in U.S. Burden of *Clostridioides difficile* infection and outcomes. *N Engl J Med* 2020;382:1320–1330.
14. Rao K, Micic D, Natarajan M, Winters S, Kiel MJ, *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;61:233–241.
15. Gonzales-Luna AJ, Carlson TJ, Dotson KM, Poblete K, Costa G, *et al*. PCR ribotypes of *Clostridioides difficile* across Texas from 2011 to 2018 including emergence of ribotype 255. *Emerg Microbes Infect* 2020;9:341–347.
16. Kamboj M, McMillen T, Syed M, Chow HY, Jani K, *et al*. Evaluation of a combined multilocus sequence typing and whole-genome sequencing two-step algorithm for routine typing of *Clostridioides difficile*. *J Clin Microbiol* 2020;59:e01955-20.
17. Endres BT, Begum K, Sun H, Walk ST, Memariani A, *et al*. Epidemic *Clostridioides difficile* ribotype 027 lineages: comparisons of Texas versus worldwide strains. *Open Forum Infect Dis* 2019;6:ofz013.
18. Martinson JNV, Broadaway S, Lohman E, Johnson C, Alam MJ, *et al*. Evaluation of portability and cost of a fluorescent PCR ribotyping protocol for *Clostridium difficile* epidemiology. *J Clin Microbiol* 2015;53:1192–1197.
19. Griffiths D, Fawley W, Kachrimanidou M, Bowden R, Crook DW, *et al*. Multilocus sequence typing of *Clostridium difficile*. *J Clin Microbiol* 2010;48:770–778.
20. Jolley KA, Bray JE, Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. *Wellcome Open Res* 2018;3:124.
21. Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, *et al*. ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads. *Microb Genomics* 2017;3:e000131.
22. Han JH, Lapp Z, Bushman F, Lautenbach E, Goldstein EJC, *et al*. Whole-genome sequencing to identify drivers of carbapenem-resistant *Klebsiella pneumoniae* transmission within and between regional long-term acute-care hospitals. *Antimicrob Agents Chemother* 2019;63:e01622-19.
23. Bolger AM, Lohse M, Usadel B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* 2014;30:2114–2120.
24. He M, Sebaihia M, Lawley TD, Stabler RA, Dawson LF, *et al*. Evolutionary dynamics of *Clostridium difficile* over short and long time scales. *Proc Natl Acad Sci U S A* 2010;107:7527–7532.
25. Yin C, Chen DS, Zhuge J, McKenna D, Sagurton 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:e01537-17.
26. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009;25:1754–1760.
27. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, *et al*. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009;25:2078–2079.
28. Croucher NJ, Page AJ, Connor TR, Delaney AJ, Keane JA, *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. *Nucleic Acids Res* 2015;43:e15.
29. 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;32:268–274.
30. Sebaihia M, Wren BW, Mullany P, Fairweather NF, Minton N, *et al*. The multidrug-resistant human pathogen *Clostridium difficile* has a highly mobile, mosaic genome. *Nat Genet* 2006;38:779–786.
31. Spigaglia P, Barbanti F, Mastrantonio P, Brazier JS, Barbut F, *et al*. Fluoroquinolone resistance in *Clostridium difficile* isolates from a prospective study of *C. difficile* infections in Europe. *J Med Microbiol* 2008;57:784–789.
32. He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, *et al*. Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*. *Nat Genet* 2013;45:109–113.
33. Popovich KJ, Snitkin ES, Hota B, Green SJ, Pirani A, *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.
34. Suchard MA, Lemey P, Baele G, Ayres DL, Drummond AJ, *et al*. Bayesian phylogenetic and phylodynamic data integration using BEAST 1.10. *Virus Evol* 2018;4:vey016.
35. Eyre DW, Didelot X, Buckley AM, Freeman J, Moura IB, *et al*. *Clostridium difficile* trehalose metabolism variants are common and not associated with adverse patient outcomes when variably present in the same lineage. *EBioMedicine* 2019;43:347–355.
36. 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:vew007.
37. Didelot X, Croucher NJ, Bentley SD, Harris SR, Wilson DJ. Bayesian inference of ancestral dates on bacterial phylogenetic trees. *Nucleic Acids Res* 2018;46:e134.
38. Duchêne S, Duchêne D, Holmes EC, SYW H. The performance of the date-randomization test in phylogenetic analyses of time-structured virus data. *Mol Biol Evol* 2015;32:1895–1906.
39. Didelot X, Eyre DW, Cule M, CL I, Ansari MA, *et al*. Microevolutionary analysis of *Clostridium difficile* genomes to investigate transmission. *Genome Biol* 2012;13:R118.
40. Minin VN, Bloomquist EW, Suchard MA. Smooth skyride through a rough skyline: Bayesian coalescent-based inference of population dynamics. *Mol Biol Evol* 2008;25:1459–1471.
41. Rambaut A, Suchard M, Xie D, Drummond A. Tracer v1.6. 2014. http://beast.bio.ed.ac.uk/Tracer
42. Loo VG, Oughton M, Bourgault AM, Kelly M, Dewar K. A predominantly clonal multi-institutional outbreak of *Clostridium difficile*–associated diarrhea with high morbidity and mortality. *N Engl J Med* 2005;8.
43. Alam MJ, Walk ST, Endres BT, Basseres E, Khaleduzzaman M, *et al*. Community environmental contamination of toxigenic *Clostridium difficile*. *Open Forum Infect Dis* 2017;4:ofx018.
44. Dingle KE, Didelot X, Quan TP, Eyre DW, Stoesser N, *et al*. Effects of control interventions on *Clostridium difficile* infection in England: an observational study. *Lancet Infect Dis* 2017;17:411–421.
45. Davies KA, Ashwin H, Longshaw CM, Burns DA, Davis GL, *et al*. Diversity of *Clostridium difficile* PCR ribotypes in Europe: results from the European, multicentre, prospective, biannual, point-prevalence study of *Clostridium difficile* infection in hospitalised patients with diarrhoea (EUCLID), 2012 and 2. *Euro Surveill Bull Eur Sur Mal Transm Eur Commun Dis Bull* 2016;21:pii30294
46. Tenover FC, Tickler IA, Persing DH. Antimicrobial-resistant strains of *Clostridium difficile* from North America. *Antimicrob Agents Chemother* 2012;56:2929–2932.
47. Foster NF, Collins DA, Ditchburn SL, Duncan CN, van Schalkwyk JW, *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;2:64–71.

48. Aitken SL, Alam MJ, Khaleduzzaman M, Khaleduzzuman M, Walk ST, *et al*. In the endemic setting, *Clostridium difficile* ribotype 027 is virulent but not hypervirulent. *Infect Control Hosp Epidemiol* 2015;36:1318–1323.
49. 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* 2016;7:2138.
50. Janezic S, Zidaric V, Pardon B, Indra A, Kokotovic B, *et al*. International *Clostridium difficile* animal strain collection and large diversity of animal associated strains. *BMC Microbiol* 2014;14:173.
51. Romano V, Pasquale V, Krovacek K, Mauri F, Demarta A, *et al*. Toxigenic *Clostridium difficile* PCR ribotypes from wastewater treatment plants in southern Switzerland. *Appl Environ Microbiol* 2012;78:6643–6646.
52. Croucher NJ, Harris SR, Grad YH, Hanage WP. Bacterial genomes in epidemiology–present and future. *Philos Trans R Soc Lond B Biol Sci* 2013;368:20120202.
53. Peacock SJ, Parkhill J, Brown NM. Changing the paradigm for hospital outbreak detection by leading with genomic surveillance of nosocomial pathogens. *Microbiology (Reading)* 2018;164:1213–1219.
54. Weller C, Wu M. A generation-time effect on the rate of molecular evolution in bacteria. *Evolution* 2015;69:643–652.
55. Carlson PE, Walk ST, Bourgis AET, Liu MW, Kopliku F, *et al*. The relationship between phenotype, ribotype, and clinical disease in human *Clostridium difficile* isolates. *Anaerobe* 2013;24:109–116.