Genomic surveillance of *Clostridioides difficile* transmission and virulence in a healthcare setting

Erin P. Newcomer[1,2*], Skye R. S. Fishbein[1,3,*], Kailun Zhang[1,3], Tiffany Hink[4], Kimberly A. Reske[4], Candice Cass[4], Zainab H. Iqbal[4], Emily L. Struttmann[4], Erik R. Dubberke[4**], Gautam Dantas[1,2,3,5,6**]

[1]The Edison Family Center for Genome Sciences and Systems Biology, Washington University School of Medicine, St. Louis, Missouri, USA.

[2]Department of Biomedical Engineering, Washington University in St. Louis, St. Louis, Missouri, USA

[3]Department of Pathology and Immunology, Division of Laboratory and Genomic Medicine, Washington University School of Medicine, St. Louis, Missouri, USA.

[4]Division of Infectious Diseases; Washington University School of Medicine, St. Louis, Missouri, USA

[5]Department of Molecular Microbiology; Washington University School of Medicine, St. Louis, Missouri, USA

[6]Department of Pediatrics, Washington University School of Medicine, St. Louis, Missouri, USA

*These authors contributed equally to this work

**Corresponding authors: ERD: edubberk@wustl.edu; GD: dantas@wustl.edu

Abstract

Clostridioides difficile infection (CDI) is a major cause of healthcare-associated diarrhea, despite the widespread implementation of contact precautions for patients with CDI. Here, we investigate strain contamination in a hospital setting and genomic determinants of disease outcomes. Across two wards over six months, we selectively cultured *C. difficile* from patients (n=384) and their environments. Whole-genome sequencing (WGS) of 146 isolates revealed that most *C. difficile* isolates were from clade 1 (131/146, 89.7%), while only one isolate of the hypervirulent ST1 was recovered. Of culture-positive admissions, 17% of patients were diagnosed with CDI upon admission. We defined 29 strain networks at ≤ 2 core gene SNPs; 2 of these networks contain strains from different patients. Strain networks were temporally linked ($p<0.0001$). Across networks and over time, we found a minority of networks contained differences in phage populations. To understand genomic correlates of disease, we conducted WGS on an additional cohort of *C. difficile* (n=102 isolates) from the same hospital and confirmed that clade 1 isolates are responsible for most CDI cases. We found that while toxigenic *C. difficile* isolates are associated with the presence of *cdtR*, nontoxigenic isolates have an increased abundance of prophages. Our pangenomic analysis of clade 1 isolates suggests that while toxin genes (*tcdABER and cdtR*) were associated with CDI symptoms, they are dispensable for patient colonization. These data indicate toxigenic and nontoxigenic *C. difficile* contamination persists in a hospital setting and highlight further investigation into how accessory genomic repertoires contribute to *C. difficile* colonization and disease.

Background

Clostridioides difficile infection (CDI) is one of the most common healthcare-associated infections (HAIs) in the US and is the leading cause of healthcare-associated infectious diarrhea[1,2]. Since the early 2000s, *C. difficile* research has focused largely on hypervirulent strains, such as PCR ribotype 027[1,3-6], which have been responsible for hospital-associated CDI epidemics. Strains of ribotype 027 were responsible for 51% and 84% of CDI cases in the US and Canada in 2005, respectively[1,4,5]. Since then, other circulating strains have emerged as the prevalent strains causative of CDI, such as 078 and 014/020[7-9]. One report indicated that the prevalence of PCR ribotype 027 decreased from 26.2% in 2012 to 16.9% in 2016[9]. As the landscape of *C. difficile* epidemiology continues to evolve, we must update our understanding of how various strains of this pathogen evolve, spread, and cause disease.

In addition to the changing prevalence of CDI-causing *C. difficile* strains, their transmission dynamics also appear to be evolving. In the late 1980s, it became clear that patients with active CDI shed spores onto their surroundings, leading to future CDI events in the healthcare setting[1]. Because of this, patients with active CDI are placed on contact precautions to prevent transmission to susceptible patients, which has been successful in reducing rates of CDI[2,10]. Nevertheless, while epidemiological estimates indicate that 20-42% of infections may be connected to a previous infection, multiple genomic studies fail to associate a CDI case to a previous case[11-13]. This suggests other potential sources of pathogen exposure in the hospital environment. While asymptomatic carriers of *C. difficile* have not been a significant focus of infection prevention efforts, studies have shown these carriers do shed viable, toxigenic *C. difficile* to their surroundings that could cause disease[14]. Several studies have shown evidence of a reduction in CDI cases if asymptomatic carriers are put on similar contact precautions to CDI patients[15-17], but this has not been consistently found[18]. Correspondingly, it is critical to understand if *C. difficile* carriers are major contributors to new *C. difficile* acquisition or CDI manifestation in hospitalized patient populations.

C. difficile strains are categorized into five major clades and three additional cryptic clades. These clades encompass immense pangenomic diversity with many mobilizable chromosomal elements[19,20], including numerous temperate phages that have potential influences over C. difficile toxin expression, sporulation, and metabolism[21]. Two major toxin loci, not required for viability, encode large multi-unit toxins that independently augment the virulence of *C. difficile*. Epithelial destruction and CDI have largely been attributed to the presence of pathogenicity locus (PaLoc) encoding toxins TcdA and TcdB. In addition, an accessory set of toxins (CdtA and CdtB) encoded at the binary toxin locus, may worsen disease symptoms[22].

Yet, many nontoxigenic strains of *C. difficile* have been documented and are adept colonizers of the GI tract, even without the PaLoc[23]. As there has been continued debate about strain-specific virulence attributes[24-26], it is important to investigate the extent of strain-level pangenomic diversity and consequences of such diversity on host disease [27,28].

The purpose of this study was to evaluate the role of *C. difficile* strain diversity in colonization outcomes and hospital epidemiology. By sampling patients (n=384) and their environments for six months in two leukemia and hematopoietic stem cell (HCT) transplant wards at Barnes-Jewish Hospital in St. Louis, USA, we used isolate genomics to identify environmental contamination of both toxigenic (TCD) and nontoxigenic (NTCD) *C. difficile* by carriers and CDI patients, and corresponding transmission between both patient groups. Longitudinal strain tracking within these transmission networks revealed accessory gene flux of multi-drug resistance loci over the course of the study. Lastly, integration of isolate genomic data and CDI information from this prospective study with isolate genomic data from a complementary retrospective study of asymptomatic vs symptomatic *C. difficile* colonization in the same hospital[29,30] indicated that the clade 1 lineage, containing both toxigenic strains and nontoxigenic strains, dominates circulating populations of *C. difficile* in this hospital. Further, this lineage of *C. difficile* has significant variation in the PaLoc operon, and harbors other genetic factors that are associated with CDI symptoms in patients.

Methods

Study Design

This prospective observational study took place in the leukemia and hematopoietic stem cell transplant (HCT) wards at Barnes-Jewish Hospital (BJH) in St. Louis, Missouri, United States. Each ward consisted of two wings with 16 beds; on the acute leukemia ward we enrolled from both wings (32 beds) and on the HCT ward we enrolled on one wing (16 beds). The wards were sampled for 6 months from January 2019-July 2019 (acute leukemia) and 4 months from March 2019-July 2019 (HCT). These units are located 2 floors apart in the same building.

Sample collection, selective culture, and isolate identification

Patients and their environments were sampled upon admission to a study ward and then weekly until discharge. Per hospital standards, bleach is used for daily and terminal discharge cleaning. From each patient, a stool specimen and/or rectal swab was collected as available. Remnant fecal samples from the BJH microbiology laboratory that were obtained during routine clinical care were also collected. Stool samples and rectal swabs collected on enrollment were refrigerated for up to 3 hours before processing. Specimens from all other timepoints were

stored in at -80°C in tryptic soy broth (TSB)/glycerol before processing. Environmental samples were collected from bedrails, keyboards, and sink surfaces using 3 E-swabs (Copan). If a surface was unable to be sampled, a swab was taken from the IV pump or nurse call button as an alternative. Swab eluate were stored at -80°C until processing.

Broth enrichment culture for *C. difficile* in Cycloserine Cefoxitin Mannitol Broth with Taurocholate and Lysozyme (CCMB-TAL) was performed on all admission specimens and checked for growth at 24h, 48h, and 7 days after inoculation. If that culture produced *C. difficile*, all other specimens collected from that patient and their surroundings were also cultured on Cycloserine-Cefoxitin Fructose Agar with Horse Blood and Taurocholate (CCFA-HT) agar. Colonies resembling *C. difficile* (large, spreading, grey, ground glass appearance) were picked by a trained microbiologist and sub-cultured onto a blood agar plate (BAP). Growth from the subculture plate was identified using Matrix-assisted laser desorption/ionization-time of flight mass spectrometry (MALDI-TOF MS). Upon identification, sweeps of *C. difficile* BAPs were collected in tryptic soy broth (TSB) and stored at -80C for sequencing. If both rectal swab sample and stool sample produced a *C. difficile* isolate, the stool isolate was preferentially used for analysis over the rectal swab isolate.

C. difficile toxin enzyme immunoassay (EIA) was conducted as part of routine clinical care based on clinical suspicion of CDI. To be diagnosed with *C. difficile* infection (CDI), a patient must have been EIA+ for *C. difficile* toxin (Alere TOX A/B II); those who weren't tested (due to no clinically significant diarrhea) or tested EIA- and were culture-positive for *C. difficile* were considered *C. difficile* carriers. Episodes of carriage or CDI are defined as the time from the first culture-positive specimen from a patient to the last culture-positive specimen during a given hospital admission.

Short read sequencing and *de novo* genome assembly

Parameters used for computational tools are provided parenthetically. Total genomic DNA from *C. difficile* isolates was extracted from frozen plate scrapes using the QIAamp BiOstic Bacteremia DNA Kit (Qiagen) and quantified DNA with the PicoGreen dsDNA assay (Thermo Fisher Scientific). DNA from each isolate was diluted to a concentration of 0.5 ng/µL for library preparation using a modified Nextera kit (Illumina) protocol[31]. Sequencing libraries were pooled and sequenced on the NovaSeq 6000 platform (Illumina) to obtain 2□×□150□bp reads. Raw reads were demultiplexed by index pair and adapter sequencing trimmed and quality filtered using Trimmomatic (v0.38, SLIDINGWINDOW:4:20, LEADING:10, TRAILING:10,

MINLEN:60)[32]. Cleaned reads were assembled into draft genomes using Unicycler (v0.4.7)[33]. Draft genome quality was assessed using Quast[34], BBMap[35], and CheckM[36], and genomes were accepted if they met the following quality standards: completeness greater than 90%, contamination less than 5%, N50 greater than 10,000 bp, and less than 500 contigs >1000bp.

Isolate characterization and typing

A Mash Screen was used to identify likely related genomes from all NCBI reference genomes[37]. Average nucleotide identity(ANI) between the top three hits and the draft assembly was calculated using dnadiff[38]. Species were determined if an isolate had >75% alignment and >96% ANI[39] to a type strain, and were otherwise classified as genomospecies of the genus level taxonomy call.

In silico multilocus sequence typing (MLST) was determined for all *C. difficile* and genomospecies isolates using mlst[40,41]. Isolate contigs were annotated using Prokka[42] (v1.14.5, -mincontiglen 500, -force, -rnammer, -proteins GCF_000210435.1_ASM21043v1_protein.faa[43]). *cdtAB* was determined to be a pseudogene if there were three hits to *cdtB*, indicating the damaged structure of the pseudogene[44]. *C. difficile* clade was determined using predefined clade-MLST relationships described in Knight, et al[19].

Phylogenetic analyses

The .gff files output by Prokka[42] were used as input for Panaroo (v1.2.10)[45] to construct a core genome alignment. The Panaroo alignment was used as input to construct a maximum-likelihood phylogenetic tree using Fasttree[46]. The output .newick file was visualized using the ggtree (v3.4.0)[47] package in R. Cryptic clade isolates were determined as such based on phylogenetic clustering with cryptic clade reference isolates.

Core genome SNP analyses and network formation

We constructed a core gene alignment for each clade using Panaroo (v1.2.10) and calling MAFFT (v7.481). We then used Gubbins (v3.3.0) to identify recombination-filtered polymorphic sites, and constructed a recombination-free polymorphic site alignment using snp-sites (v2.4.0) 25414349}[48]. We finally extracted pairwise, recombination-filtered clade specific core-gene SNP distances using snp-dists (v0.8.2)(https://github.com/tseemann/snp-dists). Strain networks were determined by connecting isolates that were <=2 SNPs from one another.

Phage identification and clustering

Isolate genomes were piped into Cenote-Taker 2[49] to identify contigs with end features as direct terminal repeats (DTRs) indicating circularity and inverted linear repeats (ITRs) or no features for linear sequences. Identified contigs were filtered by length and completeness to remove false positives. Length limits were 1,000 nucleotides (nt) for the detection of circularity, 4,000 nt for ITRs, and 5,000 nt for other linear sequences. The completeness was computed as a ratio between the length of our phage sequence and the length of matched reference genomes by CheckV[50] and the threshold was set to 10.0%. Phage contigs passing these two filters were then run through VIBRANT[51] with a "virome" flag to further remove obvious non-viral sequences[51]. Based on MIUViG recommended parameters[52], phages were grouped into "populations" if they shared ≥95% nucleotide identity across ≥85% of the genome using BLASTN and a CheckV supporting code.

Analysis of genotypic associations with disease severity

Two previously sequenced retrospective cohorts from the same hospital were included to increase statistical power[29,53]. In the analyses of toxigenic vs. nontoxigenic isolates from clade 1, Pyseer[54] was run using a SNP distance matrix (using snp-dist as above), binary genotypes(presence or absence of *tcdB*), and Panaroo-derived gene presence/absence data. In the analysis of CDI suspicion, all isolates from clade 1 were used that represented one isolate per patient-episode. Isolates recovered from environmental surfaces were excluded. Using these assemblies, a core genome alignment was generated using Prokka[42] and Panaroo[45] as above. SNP distances were inferred from the core-gene alignment using snp-dists[55]. Binary phenotypes were coded for the variable CDI suspicion, whereby isolates associated with a clinically tested stool were associated with symptomatic colonization (TRUE). Isolates that were associated with a surveillance stool and had no clinical testing associated with that patient timepoint were coded as non-symptomatic colonization (FALSE). Gene candidates filtered based on 'high-bse', and were annotated HMMER on RefSeq databases and using a bacteriophage-specific tool VIBRANT[51]. Selected outputs were visualized in R using the beta coefficient as the x-axis and the $-\log_{10}$(likelihood ratio test p-value) as the y-axis.

Reference assembly collection

We chose 23 reference assemblies from Knight, et al[19] for Figure 2c because of their MLST-clade associations (Supplementary Table 2). References span Clades 1-5 and cryptic clades C-1, C-2, and C-3, with one reference from each of the three most frequent MLSTs in each clade. Cryptic clade C-3 only had 2 reference assemblies available. References were annotated and included in phylogenetic tree construction as above.

All *Clostridioides difficile* genomes available on the National Institutes of Health (NIH) National Library of Medicine (NLM) were acquired for Figure 5c construction. References from NCBI (Supplementary Table 4) were included if they had less than 200 contigs. Assemblies that met these quality requirements were annotated and phylogenetically clustered as above.

Results

Surveillance of *C. difficile* reservoirs in hospital wards reveals patient colonization and environmental contamination.

We prospectively collected patient and environmental samples to investigate genomic determinants of *C. difficile* carriage, transmission, and CDI (Figure 1). Across the study period, we enrolled 384 patients from 654 unique hospital admissions, and collected patient specimens upon admission and weekly thereafter (Supplementary Figure 1). We collected at least one specimen (clinical stool collected as part of routine care, study collected stool, or study collected rectal swab) from 364 admissions (94.8% of enrolled patients), for a total of 1244 patient specimens. We selectively cultured *C. difficile* from 43 rectal swabs and 108 stool samples, for a total of 151 culture-positive patient specimens. We also collected weekly swabs from the bedrails, sink surfaces, and in-room keyboards, for a total of 3045 swabs from each site. In total, 22/398 (5.5%) of bedrail swabs cultured and 4/ 399 (1.0%) of keyboard swabs cultured were culture-positive for *C. difficile* (Figure 2a). *C. difficile* was never recovered from sink surfaces (all sinks on these units are hands-less activated) or other sampled sites. Collapsing multiple positive samples from the same patient admission results in 20 positive bedrails (20/79, 25.3% of all admissions with positive patient specimens) and 4 positive keyboards (4/79, 5.06% of all admissions with positive patient specimens) (Figure 2b).

C. difficile carriers outnumbered patients with CDI

Patients with CDI were identified through routine clinical care, with CDI defined as patients who had stool submitted for *C. difficile* testing, as ordered by the clinical team when suspicious for CDI, and who tested positive for *C. difficile* toxins by enzyme immunoassay (EIA+). Otherwise, if they were culture positive and EIA- or culture positive and not EIA tested, they were considered carriers. Results from selective culture indicated that 21.7% of unique admissions (79/364 admissions with available specimens) were culture-positive for *C. difficile* at some point during their admission (Figure 2b). Of culture-positive admissions, 17% (13/79) were EIA+ and diagnosed with CDI (13/364, 3.6% of all admissions with specimens available). The remaining 83% (66/79 admissions with specimens available) of culture-positive admissions

were termed carriers (Figure 2b). An additional nine admissions became EIA+ at some point during their stay for a total of 22 CDI cases, but seven did not have specimens available for culture and two were culture negative. The substantial detection of longitudinal patient *C. difficile* colonization prompted us to investigate the genomic correlates of *C. difficile*-associated disease and transmission in these two patient populations.

Phylogenetic clustering reveals lack of hypervirulent strains, presence of cryptic clades

We conducted whole-genome sequencing to ascertain phylogenetic distances among isolates and to identify closely related strains of *C. difficile.* We identified 141 isolate genomes as *C. difficile* (using a 75% alignment and 96% average nucleotide identity [ANI] threshold). One isolate was identified as *Clostridium innocuum* and five isolates were classified as *C. difficile* genomospecies (92-93% ANI). To contextualize population structure, we applied a previously established MLST-derived clade definition to our isolate cohort[19]. The majority of *C. difficile* isolates were from Clade 1 (131/146, 89.7% of *C. difficile* and genomospecies, Figure 2c). Four patient-derived isolates were identified from clade 2, but only one was of the hypervirulent strain ST1 (PCR ribotype 027)[6]. We found that the distribution of STs associated with carriers was significantly different from that of STs associated with CDI patients ($p<0.001$, Fisher's exact test) suggesting some strain-specificity to disease outcome.

Interestingly, the five genomospecies isolates clustered with other isolates belonging to a recently discovered *C. difficile* cryptic clade C-1 (Supplementary Figure 2). While cryptic clades are genomically divergent from *C. difficile*, these isolates can produce homologs to TcdA/B and cause CDI-like disease in humans[19,56]. In a clinical setting, they are frequently identified by MALDI-TOF MS as *C. difficile* and diagnosed as causative of CDI[56]. These data highlight the novel distribution of circulating *C. difficile* strains in the two study wards. While many patients with multiple isolates had homogeneous signatures of colonization (with closely related isolates), four patients (4/72, 6%) produced isolates from distinct ST types.

Carriers and CDI patients contribute to transmission networks and environmental contamination

Given the predominance of Clade 1 isolates, we sought to identify clonal populations of *C. difficile* strains, indicative of direct *C. difficile* contamination (patient-environment) or transmission (patient-patient). We compared pairwise, recombination-filtered within-clade core gene single nucleotide polymorphism (SNP) distances to identify networks of transmission connecting isolates <=2 SNPs apart (Supplementary Figure 4). We identified a total of 29 strain networks, 2 of which contain patient isolates from different patients (Figure 3a). These strain

networks were temporally linked, as there were significantly fewer days between same-network isolates than isolates from different networks ($p<2.2e-16$, Wilcoxon, Figure 3b). We compared strain connections among a single patient's isolates from stool or rectal swab('patient'), and between these isolates and environmental isolates from their immediate surroundings ('bedrail' or 'keyboard', Figure 3c). While the majority of bedrail isolates fell within the same network as patient isolates from that room (30 of 44 comparisons, 68%), 32% (14 of 44 comparisons) were genomically distinct, suggesting contamination from alternate sources. Keyboards were mostly colonized with distinct strains from the patient (22%, 2/9 comparisons were the same strain), indicating other routes of contamination ($p<0.05$, Fisher's exact test, BH corrected. Figure 3c). Among the networks that contain multiple patients, we found no instances of potential transmission from the inhabitant of one room to the subsequent inhabitant. However, in both instances, each potential transmission was associated with a temporal overlap in patient stay in the same ward, providing epidemiological capacity for transmission ($p<0.05$, Wilcoxon test). Importantly, we found no networks connecting patients with CDI to *C. difficile* carriers, suggesting successful containment through contact precaution protocols. These data highlight multiple sources of environmental contamination by *C. difficile* and prompted us to investigate the relationship between genetic factors and patient symptomology.

Phage populations persist in circulating *C.difficile* networks

C. difficile isolates have an extensive pangenome, with genetic loci mobilized by conjugative elements and phages, and mobilizable elements playing a key role in *C. difficile*'s lifecycle[57]. Temperate phages, which can undergo lytic replication or insert into the host genome as a latent prophage, are the only phages that have been isolated for *C. difficile*[58]. To identify *C. difficile* prophage signatures and understand how dynamic they were in our strain networks, we analyzed our isolate genomes with Cenote-Taker 2 for putative phage contigs. After filtering for quality, we grouped contigs into phage populations (vOTUs) and quantified the alpha-diversity of phage populations in each isolate, and across MLST types (Figure 4a). ST42 and ST2, some of the most globally abundant ST types had the lowest diversity of phages in our cohort, though this negative correlation was not statistically significant across ST types (Figure 4b; $R=-0.31$, $p=0.12$). Our clonality-resolved strain networks allowed us to investigate phage flux over time. We found that the majority of networks (23/29) carried the same number of phages over time (Figure 4c), suggesting persistent roles in *C. difficile* biology. Interestingly, we found that nontoxigenic isolates had a higher diversity of phage populations relative to toxigenic isolates

(Figure 4d). These data suggest distinct selective pressures on temperate phages in *C. difficile* related to toxin gene presence.

Accessory genomic elements are associated with host CDI symptoms

Despite evidence of transmission in this prospective study, a minority of patients were diagnosed with CDI relative to those asymptomatically colonized with *C. difficile* in part due to the presence of nontoxigenic *C. difficile* isolates (Figure 2b). To power our investigation of virulence determinants across patient-colonizing *C. difficile* strains, we performed whole genome sequencing on 102 additional patient-derived *C. difficile* isolates from a previously described *C. difficile*-colonized/CDI cohort from the same hospital[29], where all patients had clinical suspicion of CDI (CDI suspicion), defined by a clinician ordering an EIA test during patient admission. Using an MLST-based clade definition as above, we identified that most CDI cases result from isolates within clade 1, though clade 2 isolates were more likely to be associated with CDI status (Figure 5a). The latter finding supports previous data indicating that clade 2 isolates are hypervirulent, often attributed to the presence of the binary toxin operon or increased expression from the PaLoc[22,59,60]. Meanwhile, some clade 1 isolates contain no toxins, indicating a diversity of colonization strategies in this lineage. Pangenomic comparison of nontoxigenic versus toxigenic isolates revealed that in addition to the PaLoc, the majority of our toxigenic isolates from clade 1 (95/131 of our cohort) possess remnants of the binary toxin operon (Figure 5b, *cdtR and cdtA/B* pseudogenes). Given the previous report that full-length *cdtAB* was identified only within Clades 2, 3, and 5[19], we investigated the conservation of *cdtR* (the transcriptional regulator of the binary toxin locus) across *C. difficile* strains (containing 5 lineages). We additionally examined >1400 *C. difficile* genome assemblies from NCBI (Supplementary Table 4, Figure 5c). *cdtR* (unlike *cdtAB*) was dispersed across clade 1 and significantly associated with *tcdB* (Figure 5d, Fisher's exact test, BH corrected), suggesting a selective pressure to maintain some element of both toxin loci in these isolates. Notably, these operons are not syntenic, further underlining the significance of the association. From this association, we sought to further understand why some toxigenic clade 1 isolates cause CDI and some colonize without symptoms. Using 148 toxigenic clade 1 isolates collected from this study and two previous studies from the same hospital[29,53], we utilized a bacterial GWAS approach, *pyseer*[54], that identifies genetic traits associated with strains corresponding to patients with CDI symptoms. Using CDI suspicion (see Methods) as an outcome variable, we found that, multiple amidases (including *cwlD*), putative transcriptional regulators, and many genes of unknown function were enriched in isolates associated with CDI symptoms (Figure

5e). These data indicate that the most prevalent, circulating *Cd* strains that cause CDI are not the hypervirulent clade 2 strains, but highlight the possibility that remnant genomic features from epidemic strains and other features may contribute to virulence in this hospital clade of *C. difficile*.

Discussion

Through our prospective genomics study of two hospital wards, we were able to identify connections between the contamination of different surfaces and the strains carried by hospitalized patients and quantify some spread between carriers. Our estimates of the prevalence of patients with CDI (3.8%) agree with other estimates of 2-4% CDI in patients with cancer[61-63]. While many studies have quantified surface contamination, few have had the genomic resolution to identify clonality between isolates indicating transmission or patient shedding[64-66]. We observed distinct patterns of contamination between a patient's bedrail and the corresponding room keyboard, supporting the notion that the bedrail could be one of multiple critical points of transmission in a hospital setting. Further, we did not identify any instances of CDI that could be genomically linked to an earlier CDI case or *C. difficile* carrier. Despite the small sample size, these data support the continued use of contact precautions for CDI patients[18].

Our data suggests the need to continually update our understanding of CDI-causing *C. difficile* strains beyond previous epidemic strains to clarify mechanisms of how the most prevalent strains relate to transmission and disease. Across 146 patient specimens, we only identified one incidence of the epidemic ST1 strain. This ribotype caused one case of CDI within our cohort, corroborating the decline in this epidemic lineage[67]. Because the overall burden of Clade 1 isolates was so high, we hypothesize that its ability to colonize without causing CDI could allow for a substantial expansion of transmission networks (especially for the case of nontoxigenic strains). While Clade 1 isolates associated with CDI symptoms are expectedly toxigenic (containing the toxin genes in the PaLoc), we also found an enrichment in two different amidase genes, that could either contribution to differences in germination rate or possess endolysin function[68 69]. How the function of such a gene contributes to an increase in symptomology remains to be understood. Further, we confirmed a genetic relationship between *cdtR* and *tcdB* across *C. difficile* lineages that indicates some evolutionary pressure for maintaining the regulatory gene of the less prevalent toxin operon (*cdtR*). This phylogenomic analysis supports recent functional data from clade 2 isolates, where the presence of full-length *cdtR* increases the expression of *tcdB* and disease severity in an animal model of CDI[57]. While

this was previously suggested *in vitro*, it is unclear how generalizable this relationship is across lineages[59]. In fact, we predict that clade 1 isolates containing only *cdtR* and the PaLoc may produce more toxin *in vivo*. Future studies are warranted to investigate the role of both classes of genes implicated in this phenotype.

Our study contextualizes the need for investigating *C. difficile* evolution within patients over time, especially concerning functional mobile units such as temperate phages. We examined phage populations in our isolates as they are a relevant mobile unit of the *C. difficile* pangenome and their stability over time has not been systematically investigated. While we find that the majority of *C. difficile* strains maintain their diversity of phage populations over time, we acknowledge that hospital admission is a prescribed period of time and we may be underestimating the amount that phage diversity changes in isolates over longer periods of time *in vivo*. Our quantification of increased phage diversity in nontoxigenic isolates suggests phage niche specialization based on the presence of the PaLoc. It is noteworthy that early characterization of the PaLoc operon indicated that it was integrated into the *C. difficile* chromosome by an ancient prophage[70-72]. Future work is required to understand how persistent phages function during *C. difficile* growth and pathogenesis[58].

Our study has a number of important limitations. As this study focused on *C. difficile* colonization, disease, and transmission in two wards in the same hospital, studies with increased sample size or meta-analysis studies are necessary to understand generalizable epidemiological measurements of *C. difficile*-patient dynamics[73]. Additionally, our study protocol allowed for culturing all environmental/patient specimens from a carrier or patient with CDI. Thus, It is possible that our estimate of carriage in this study population is an overestimate. Finally, we note the evidence for multi-strain colonization within a single patient (Patient 2330). Given our approach of only culturing and sequencing single isolates per patient timepoint, future studies are needed to investigate the extent of within-patient *C. difficile* strain diversity by interrogating additional cultured isolates per samples[74] or via metagenomic methods.

Despite these limitations, this work allows us to understand an updated genomic picture of circulating *C. difficile* in hospital-associated patients: how strains spread, their evolution, and their virulence potential in this study population. Indeed, though much human and animal research has focused on epidemic strains that are two decades old, we and others have identified more disease and colonization from distinct lineages of *C. difficile*, namely clade 1 lineages. Moreover, within this lineage we found a mosaic representation of the PaLoc that highlighted the possibility of different mechanisms of colonization and virulence by this

population of *C. difficile*. Future studies utilizing other human cohorts or animal models are warranted to investigate disease and pathogenicity caused by Clade 1 *C. difficile* strains.

List of Abbreviations

BAP: blood agar plate

CCFA-HT: Cycloserine-Cefoxitin Fructose Agar with Horse Blood and Taurocholate

CCMB-TAL: Cycloserine Cefoxitin Mannitol Broth with Taurocholate and Lysozyme

CDI: *Clostridioides difficile* infection

EIA: enzyme immunoassay

HAI: healthcare-associated infection

HGT: horizontal gene transfer

MALDI-TOF MS: Matrix-assisted laser desorption/ionization-time of flight mass spectrometry

NTCD: non-toxigenic *C. difficile*

PaLoc: pathogenicity locus

TCD: toxigenic *C. difficile*

TSB: tryptic soy broth

Declarations

Ethics approval and consent to participate

The study protocol was approved by the Washington University Human Research Protection Office (IRB #201810103). All participants provided written informed consent.

Consent for publication

Not applicable.

Availability of data and materials

The datasets generated and analyzed during the current study are available in NCBI GenBank under BioProject accession no. PRJNA980715.

Competing interests

The authors declare that they have no competing interests.

Funding

This work was supported in part by an award to ERD and GD through the Foundation for Barnes-Jewish Hospital and Institute of Clinical and Translational Sciences. This publication was supported by the NIH/National Center for Advancing Translational Sciences (NCATS), grant UL1 TR002345 (PI: B. Evanoff). This work was also supported by funding through the CDC BAA #200-2018-02926 under PI Erik Dubberke. SRSF is supported by the National Institute of Child Health and Human Development (NICHD: https://www.nicdhd.nih/gov) of the NIH under award number T32 HD004010 (PI: P. Tarr). The conclusions from this study represent those of the authors and do not represent positions of the funding agencies.

Authors' contributions

SRSF, KAR, ERD, and GD participated in idea formulation and funding for this project. TH, KAR, CC, ZHI, ELS, and ERD conducted participant enrollment, sample collection, and microbiological isolation. EPN, SRSF, KZ, and GD conducted all sequencing analysis and figure generation. EPN and SRSF completed the writing of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

The authors are grateful for members of the Dantas lab for their helpful feedback on the data analysis and preparation of the manuscript. The authors would also like to thank the Edison Family Center for Genome Sciences and Systems Biology staff, Eric Martin, Brian Koebbe, MariaLynn Crosby, and Jessica Hoisington-López for their expertise and support in sequencing/data analysis.

References

1 Czepiel, J. *et al.* Clostridium difficile infection: review. *Eur J Clin Microbiol Infect Dis* **38**, 1211-1221, doi:10.1007/s10096-019-03539-6 (2019).

2 McDonald, L. C. *et al.* Clinical Practice Guidelines for Clostridium difficile Infection in Adults and Children: 2017 Update by the Infectious Diseases Society of America (IDSA) and Society for Healthcare Epidemiology of America (SHEA). *Clin Infect Dis* **66**, e1-e48, doi:10.1093/cid/cix1085 (2018).

3 Clements, A. C., Magalhaes, R. J., Tatem, A. J., Paterson, D. L. & Riley, T. V. Clostridium difficile PCR ribotype 027: assessing the risks of further worldwide spread. *Lancet Infect Dis* **10**, 395-404, doi:10.1016/S1473-3099(10)70080-3 (2010).

4 McDonald, L. C. *et al.* An epidemic, toxin gene-variant strain of Clostridium difficile. *N Engl J Med* **353**, 2433-2441, doi:10.1056/NEJMoa051590 (2005).

5 Loo, V. G. *et al.* A predominantly clonal multi-institutional outbreak of Clostridium difficile-associated diarrhea with high morbidity and mortality. *N Engl J Med* **353**, 2442-2449, doi:10.1056/NEJMoa051639 (2005).

6 Fatima, R. & Aziz, M. The Hypervirulent Strain of Clostridium Difficile: NAP1/B1/027 - A Brief Overview. *Cureus* **11**, e3977, doi:10.7759/cureus.3977 (2019).

7 Goorhuis, A. *et al.* Emergence of Clostridium difficile infection due to a new hypervirulent strain, polymerase chain reaction ribotype 078. *Clin Infect Dis* **47**, 1162-1170, doi:10.1086/592257 (2008).

8 Bauer, M. P. *et al.* Clostridium difficile infection in Europe: a hospital-based survey. *Lancet* **377**, 63-73, doi:10.1016/S0140-6736(10)61266-4 (2011).

9 Giancola, S. E., Williams, R. J., 2nd & Gentry, C. A. Prevalence of the Clostridium difficile BI/NAP1/027 strain across the United States Veterans Health Administration. *Clin Microbiol Infect* **24**, 877-881, doi:10.1016/j.cmi.2017.11.011 (2018).

10 Balsells, E. *et al.* Infection prevention and control of Clostridium difficile: a global review of guidelines, strategies, and recommendations. *J Glob Health* **6**, 020410, doi:10.7189/jogh.06.020410 (2016).

11 Kong, L. Y. *et al.* Clostridium difficile: Investigating Transmission Patterns Between Infected and Colonized Patients Using Whole Genome Sequencing. *Clin Infect Dis* **68**, 204-209, doi:10.1093/cid/ciy457 (2019).

12 Svenungsson, B. *et al.* Epidemiology and molecular characterization of Clostridium difficile strains from patients with diarrhea: low disease incidence and evidence of limited cross-infection in a Swedish teaching hospital. *J Clin Microbiol* **41**, 4031-4037, doi:10.1128/JCM.41.9.4031-4037.2003 (2003).

13 Durham, D. P., Olsen, M. A., Dubberke, E. R., Galvani, A. P. & Townsend, J. P. Quantifying Transmission of Clostridium difficile within and outside Healthcare Settings. *Emerg Infect Dis* **22**, 608-616, doi:10.3201/eid2204.150455 (2016).

14 Warren, B. G. *et al.* The Impact of Infection Versus Colonization on Clostridioides difficile Environmental Contamination in Hospitalized Patients With Diarrhea. *Open Forum Infect Dis* **9**, ofac069, doi:10.1093/ofid/ofac069 (2022).

15 Longtin, Y. *et al.* Effect of Detecting and Isolating Clostridium difficile Carriers at Hospital Admission on the Incidence of C difficile Infections: A Quasi-Experimental Controlled Study. *JAMA Intern Med* **176**, 796-804, doi:10.1001/jamainternmed.2016.0177 (2016).

16 Xiao, Y. *et al.* Impact of Isolating Clostridium difficile Carriers on the Burden of Isolation Precautions: A Time Series Analysis. *Clin Infect Dis* **66**, 1377-1382, doi:10.1093/cid/cix1024 (2018).

17 Grigoras, C. A., Zervou, F. N., Zacharioudakis, I. M., Siettos, C. I. & Mylonakis, E. Isolation of C. difficile Carriers Alone and as Part of a Bundle Approach for the Prevention of Clostridium difficile Infection (CDI): A Mathematical Model Based on Clinical Study Data. *PLoS One* **11**, e0156577, doi:10.1371/journal.pone.0156577 (2016).

18 Morgan, D. J. *et al.* The Impact of Universal Glove and Gown Use on Clostridioides Difficile Acquisition: A Cluster-Randomized Trial. *Clin Infect Dis* **76**, e1202-e1207, doi:10.1093/cid/ciac519 (2023).

19 Knight, D. R. *et al.* Major genetic discontinuity and novel toxigenic species in Clostridioides difficile taxonomy. *Elife* **10**, doi:10.7554/eLife.64325 (2021).

20 Mullany, P., Allan, E. & Roberts, A. P. Mobile genetic elements in Clostridium difficile and their role in genome function. *Res Microbiol* **166**, 361-367, doi:10.1016/j.resmic.2014.12.005 (2015).

21 Fortier, L. C. Bacteriophages Contribute to Shaping Clostridioides (Clostridium) difficile Species. *Front Microbiol* **9**, 2033, doi:10.3389/fmicb.2018.02033 (2018).

22 Gerding, D. N., Johnson, S., Rupnik, M. & Aktories, K. Clostridium difficile binary toxin CDT: mechanism, epidemiology, and potential clinical importance. *Gut Microbes* **5**, 15-27, doi:10.4161/gmic.26854 (2014).

23 Gerding, D. N. *et al.* Administration of spores of nontoxigenic Clostridium difficile strain M3 for prevention of recurrent C. difficile infection: a randomized clinical trial. *JAMA* **313**, 1719-1727, doi:10.1001/jama.2015.3725 (2015).

24 Carlson, P. E., Jr. *et al.* The relationship between phenotype, ribotype, and clinical disease in human Clostridium difficile isolates. *Anaerobe* **24**, 109-116, doi:10.1016/j.anaerobe.2013.04.003 (2013).

25 Walk, S. T. *et al.* Clostridium difficile ribotype does not predict severe infection. *Clin Infect Dis* **55**, 1661-1668, doi:10.1093/cid/cis786 (2012).

26 Aitken, S. L. *et al.* In the Endemic Setting, Clostridium difficile Ribotype 027 Is Virulent But Not Hypervirulent. *Infect Control Hosp Epidemiol* **36**, 1318-1323, doi:10.1017/ice.2015.187 (2015).

27 Pettit, L. J. *et al.* Functional genomics reveals that Clostridium difficile Spo0A coordinates sporulation, virulence and metabolism. *BMC Genomics* **15**, 160, doi:10.1186/1471-2164-15-160 (2014).

28 Awad, M. M., Johanesen, P. A., Carter, G. P., Rose, E. & Lyras, D. Clostridium difficile virulence factors: Insights into an anaerobic spore-forming pathogen. *Gut Microbes* **5**, 579-593, doi:10.4161/19490976.2014.969632 (2014).

29 Fishbein, S. R. *et al.* Multi-omics investigation of Clostridioides difficile-colonized patients reveals pathogen and commensal correlates of C. difficile pathogenesis. *Elife* **11**, doi:10.7554/eLife.72801 (2022).

30 Dubberke, E. R. *et al.* Clostridium difficile colonization among patients with clinically significant diarrhea and no identifiable cause of diarrhea. *Infect Control Hosp Epidemiol* **39**, 1330-1333, doi:10.1017/ice.2018.225 (2018).

31 Baym, M. *et al.* Inexpensive multiplexed library preparation for megabase-sized genomes. *PLoS One* **10**, e0128036, doi:10.1371/journal.pone.0128036 (2015).

32 Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120, doi:10.1093/bioinformatics/btu170 (2014).

33 Wick, R. R., Judd, L. M., Gorrie, C. L. & Holt, K. E. Unicycler: Resolving bacterial genome assemblies from short and long sequencing reads. *PLoS Comput Biol* **13**, e1005595, doi:10.1371/journal.pcbi.1005595 (2017).

34 Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075, doi:10.1093/bioinformatics/btt086 (2013).

35 Bushnell, B. *BBMap: A Fast, Accurate, Splice-Aware Aligner*, <https://www.osti.gov/servlets/purl/1241166> (2014).

36 Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**, 1043-1055, doi:10.1101/gr.186072.114 (2015).

37 Ondov, B. D. *et al.* Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**, 132, doi:10.1186/s13059-016-0997-x (2016).

38 Kurtz, S. *et al.* Versatile and open software for comparing large genomes. *Genome Biol* **5**, R12, doi:10.1186/gb-2004-5-2-r12 (2004).

39 Richter, M. & Rossello-Mora, R. Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A* **106**, 19126-19131, doi:10.1073/pnas.0906412106 (2009).

40 Seemann, T. *mlst*, <https://github.com/tseemann/mlst> (

41 Jolley, K. A. & Maiden, M. C. BIGSdb: Scalable analysis of bacterial genome variation at the population level. *BMC Bioinformatics* **11**, 595, doi:10.1186/1471-2105-11-595 (2010).

42 Seemann, T. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* **30**, 2068-2069, doi:10.1093/bioinformatics/btu153 (2014).

43 He, M. *et al.* Evolutionary dynamics of Clostridium difficile over short and long time scales. *Proc Natl Acad Sci U S A* **107**, 7527-7532, doi:10.1073/pnas.0914322107 (2010).

44 Carter, G. P. *et al.* Binary toxin production in Clostridium difficile is regulated by CdtR, a LytTR family response regulator. *J Bacteriol* **189**, 7290-7301, doi:10.1128/JB.00731-07 (2007).

45 Tonkin-Hill, G. *et al.* Producing polished prokaryotic pangenomes with the Panaroo pipeline. *Genome Biol* **21**, 180, doi:10.1186/s13059-020-02090-4 (2020).

46 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. *Mol Biol Evol* **26**, 1641-1650, doi:10.1093/molbev/msp077 (2009).

47 Yu, G. Using ggtree to Visualize Data on Tree-Like Structures. *Curr Protoc Bioinformatics* **69**, e96, doi:10.1002/cpbi.96 (2020).

48 Page, A. J. *et al.* SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. *Microb Genom* **2**, e000056, doi:10.1099/mgen.0.000056 (2016).

49 Tisza, M. J., Belford, A. K., Dominguez-Huerta, G., Bolduc, B. & Buck, C. B. Cenote-Taker 2 democratizes virus discovery and sequence annotation. *Virus Evol* **7**, veaa100, doi:10.1093/ve/veaa100 (2021).

50 Nayfach, S. *et al.* CheckV assesses the quality and completeness of metagenome-assembled viral genomes. *Nat Biotechnol* **39**, 578-585, doi:10.1038/s41587-020-00774-7 (2021).

51 Kieft, K., Zhou, Z. & Anantharaman, K. VIBRANT: automated recovery, annotation and curation of microbial viruses, and evaluation of viral community function from genomic sequences. *Microbiome* **8**, 90, doi:10.1186/s40168-020-00867-0 (2020).

52 Roux, S. *et al.* Minimum Information about an Uncultivated Virus Genome (MIUViG). *Nat Biotechnol* **37**, 29-37, doi:10.1038/nbt.4306 (2019).

53 Fishbein, S. R. S. *et al.* Randomized Controlled Trial of Oral Vancomycin Treatment in Clostridioides difficile-Colonized Patients. *mSphere* **6**, doi:10.1128/mSphere.00936-20 (2021).

54 Lees, J. A., Galardini, M., Bentley, S. D., Weiser, J. N. & Corander, J. pyseer: a comprehensive tool for microbial pangenome-wide association studies. *Bioinformatics* **34**, 4310-4312, doi:10.1093/bioinformatics/bty539 (2018).

55 Seemann, T. *snippy: fast bacterial variant calling from NGS reads*, <https://github.com/tseemann/snippy> (2015).

56 Williamson, C. H. D. *et al.* Identification of novel, cryptic Clostridioides species isolates from environmental samples collected from diverse geographical locations. *Microb Genom* **8**, doi:10.1099/mgen.0.000742 (2022).

57 Dong, Q. *et al.* Virulence and genomic diversity among clinical isolates of ST1 (BI/NAP1/027) Clostridioides difficile. *Cell Rep* **42**, 112861, doi:10.1016/j.celrep.2023.112861 (2023).

58 Heuler, J., Fortier, L. C. & Sun, X. Clostridioides difficile phage biology and application. *FEMS Microbiol Rev* **45**, doi:10.1093/femsre/fuab012 (2021).

59 Lyon, S. A., Hutton, M. L., Rood, J. I., Cheung, J. K. & Lyras, D. CdtR Regulates TcdA and TcdB Production in Clostridium difficile. *PLoS Pathog* **12**, e1005758, doi:10.1371/journal.ppat.1005758 (2016).

60 Dong, Q. *et al.* Virulence and genomic diversity among clinical isolates of ST1 (BI/NAP1/027) Clostridioides difficile. *bioRxiv*, doi:10.1101/2023.01.12.523823 (2023).

61 Zheng, Y. *et al.* Clostridium difficile colonization in preoperative colorectal cancer patients. *Oncotarget* **8**, 11877-11886, doi:10.18632/oncotarget.14424 (2017).

62 Jain, T. *et al.* Clostridium Difficile Colonization in Hematopoietic Stem Cell Transplant Recipients: A Prospective Study of the Epidemiology and Outcomes Involving Toxigenic and Nontoxigenic Strains. *Biol Blood Marrow Transplant* **22**, 157-163, doi:10.1016/j.bbmt.2015.07.020 (2016).

63 Kamboj, M., Gennarelli, R. L., Brite, J., Sepkowitz, K. & Lipitz-Snyderman, A. Risk for Clostridiodes difficile Infection among Older Adults with Cancer. *Emerg Infect Dis* **25**, 1683-1689, doi:10.3201/eid2509.181142 (2019).

64 Claro, T., Daniels, S. & Humphreys, H. Detecting Clostridium difficile spores from inanimate surfaces of the hospital environment: which method is best? *J Clin Microbiol* **52**, 3426-3428, doi:10.1128/JCM.01011-14 (2014).

65 Kumar, N. *et al.* Genome-Based Infection Tracking Reveals Dynamics of Clostridium difficile Transmission and Disease Recurrence. *Clin Infect Dis* **62**, 746-752, doi:10.1093/cid/civ1031 (2016).

66 Kiersnowska, Z. M., Lemiech-Mirowska, E., Michalkiewicz, M., Sierocka, A. & Marczak, M. Detection and Analysis of Clostridioides difficile Spores in a Hospital Environment. *Int J Environ Res Public Health* **19**, doi:10.3390/ijerph192315670 (2022).

67 Snydman, D. R. *et al.* Epidemiologic trends in Clostridioides difficile isolate ribotypes in United States from 2011 to 2016. *Anaerobe* **63**, 102185, doi:10.1016/j.anaerobe.2020.102185 (2020).

68 Diaz, O. R., Sayer, C. V., Popham, D. L. & Shen, A. Clostridium difficile Lipoprotein GerS Is Required for Cortex Modification and Thus Spore Germination. *mSphere* **3**, doi:10.1128/mSphere.00205-18 (2018).

69 Wydau-Dematteis, S. *et al.* Cwp19 Is a Novel Lytic Transglycosylase Involved in Stationary-Phase Autolysis Resulting in Toxin Release in Clostridium difficile. *mBio* **9**, doi:10.1128/mBio.00648-18 (2018).

70 Canchaya, C., Proux, C., Fournous, G., Bruttin, A. & Brussow, H. Prophage genomics. *Microbiol Mol Biol Rev* **67**, 238-276, table of contents, doi:10.1128/MMBR.67.2.238-276.2003 (2003).

71 Proux, C. *et al.* The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. *J Bacteriol* **184**, 6026-6036, doi:10.1128/JB.184.21.6026-6036.2002 (2002).

72 Goh, S., Chang, B. J. & Riley, T. V. Effect of phage infection on toxin production by Clostridium difficile. *J Med Microbiol* **54**, 129-135, doi:10.1099/jmm.0.45821-0 (2005).

73 Miles-Jay, A. *et al.* Longitudinal genomic surveillance of carriage and transmission of Clostridioides difficile in an intensive care unit. *Nat Med*, doi:10.1038/s41591-023-02549-4 (2023).

74 Seekatz, A. M. *et al.* Presence of multiple Clostridium difficile strains at primary infection is associated with development of recurrent disease. *Anaerobe* **53**, 74-81, doi:10.1016/j.anaerobe.2018.05.017 (2018).

Figure Titles and Captions

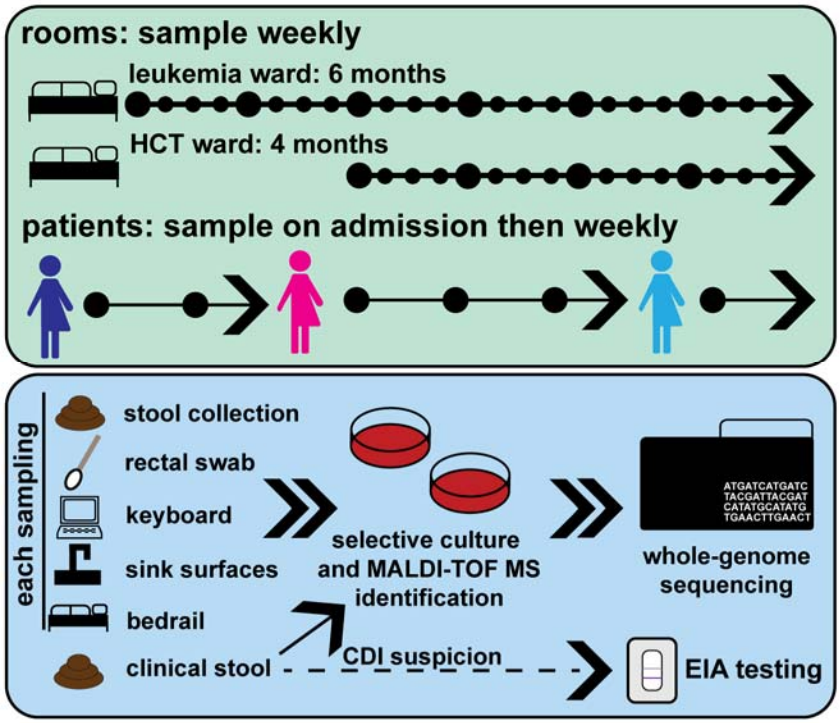


Figure 1: Study sampling and testing overview.

Caption: a) We sampled a leukemia and hematopoietic stem cell transplant ward at Barnes-Jewish Hospital in St. Louis, USA for 6 and 4 months respectively. Patients were enrolled and sampled upon admission, and then weekly for their time in the study wards. Surfaces were sampled weekly across the duration of the study. All samples and stool collected as part of routine clinical care were subjected to selective culture and MALDI-TOF MS identification, and isolates were whole-genome sequenced. Results of EIA testing as part of routine care were obtained.

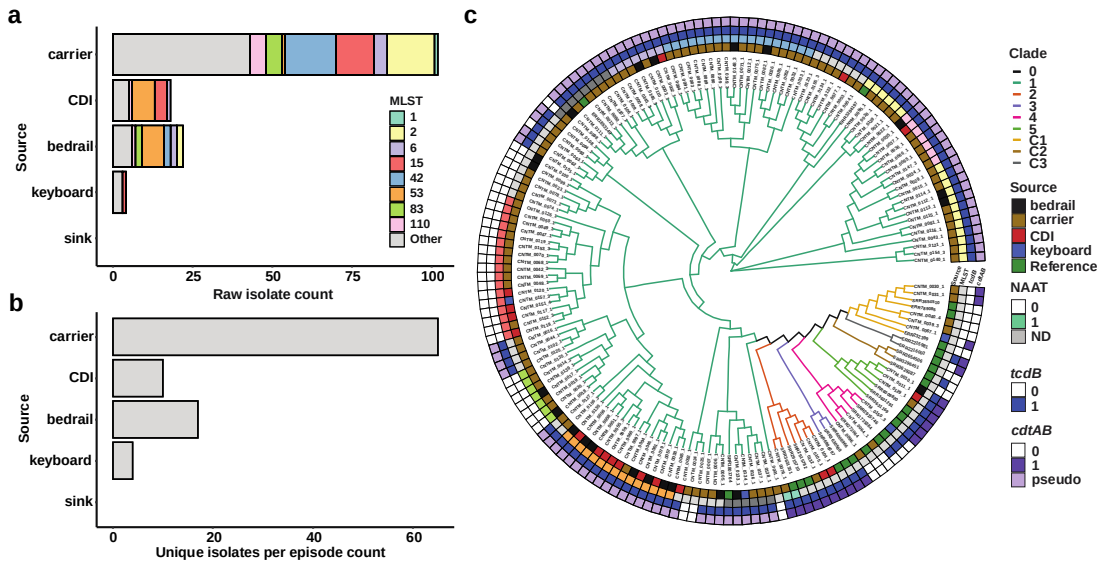


Figure 2: Total samples collected and phylogenetic relationships reveal carriers outnumber CDI patients and bedrails are the most commonly contaminated surface.

Caption: Total a) isolates collected and b) culture-positive episodes from each source. We found more carriers than CDI patients, and bedrails yielded the most *C. difficile* isolates. c) Cladogram of all isolates collected during this study plus references.

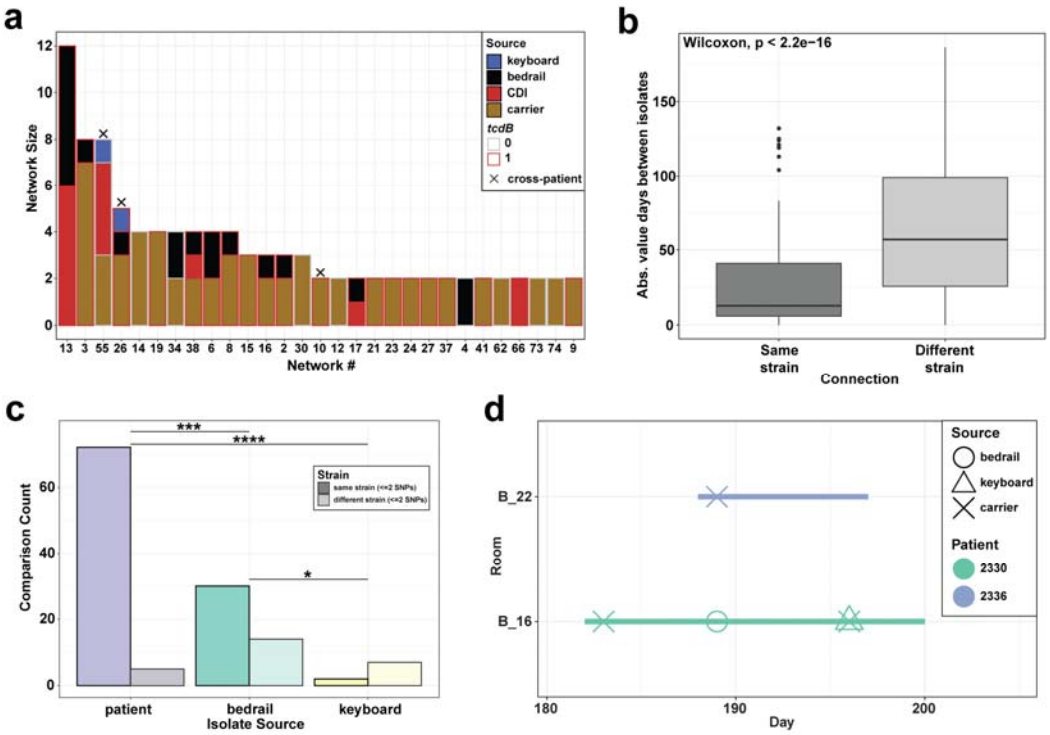


Figure 3: Hospital bedrails are a site of environmental contamination from colonized and CDI patients.

Caption: a) Strain networks were defined by <=2 core gene SNP cutoff. Network 55 includes the non-toxigenic isolates from Patient 2245 that are likely not responsible for the CDI. b) Absolute value of days between isolates within strains and between strains. Isolates within the same strain were significantly temporally linked (p<2.2e-16, Wilcoxon test). c) Number of comparisons in each group that fall within strain cutoff. Fisher's exact test, BH corrected. d) Strain tracking diagram of transmission network 26, colors indicate patients and horizontal lines indicate stay in a room. Patient 2336 sheds *C. difficile* onto the bedrail in room B_16, and patient 2330 later is identified as a carrier of the same strain.

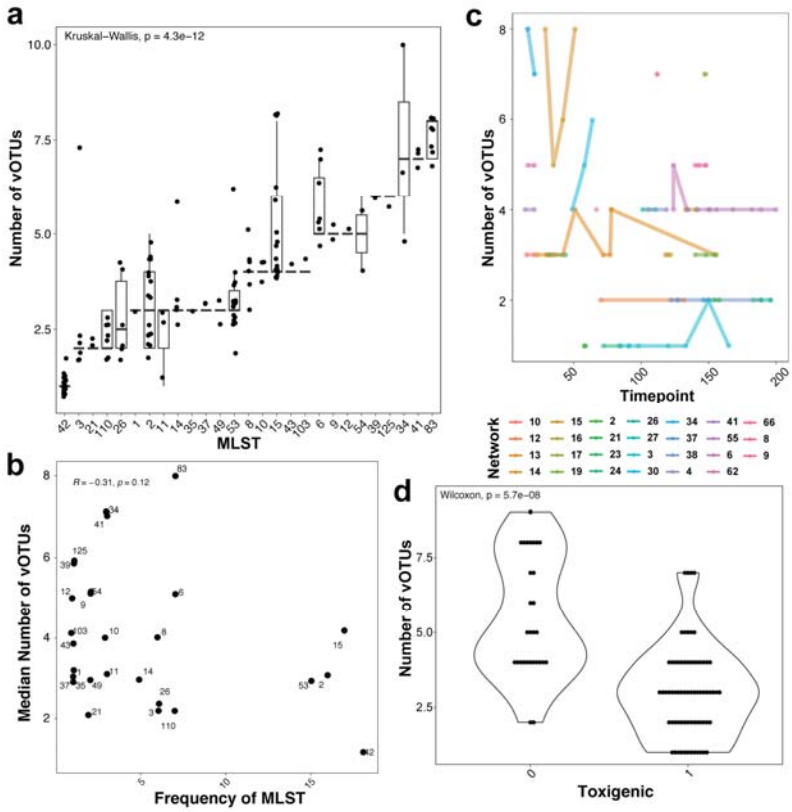


Figure 4: Phage persistence in circulating *C. difficile* networks.

Caption: a) Phage diversity measured by phage population abundance for each isolate within an MLST. b) Relationship between phage diversity and frequency of ST in our cohort c)Temporal trajectory of phage diversity for each network over time. d) phage population richness across toxigenic and nontoxigenic isolates in our cohort, Wilcoxon test, $p<0.001$.

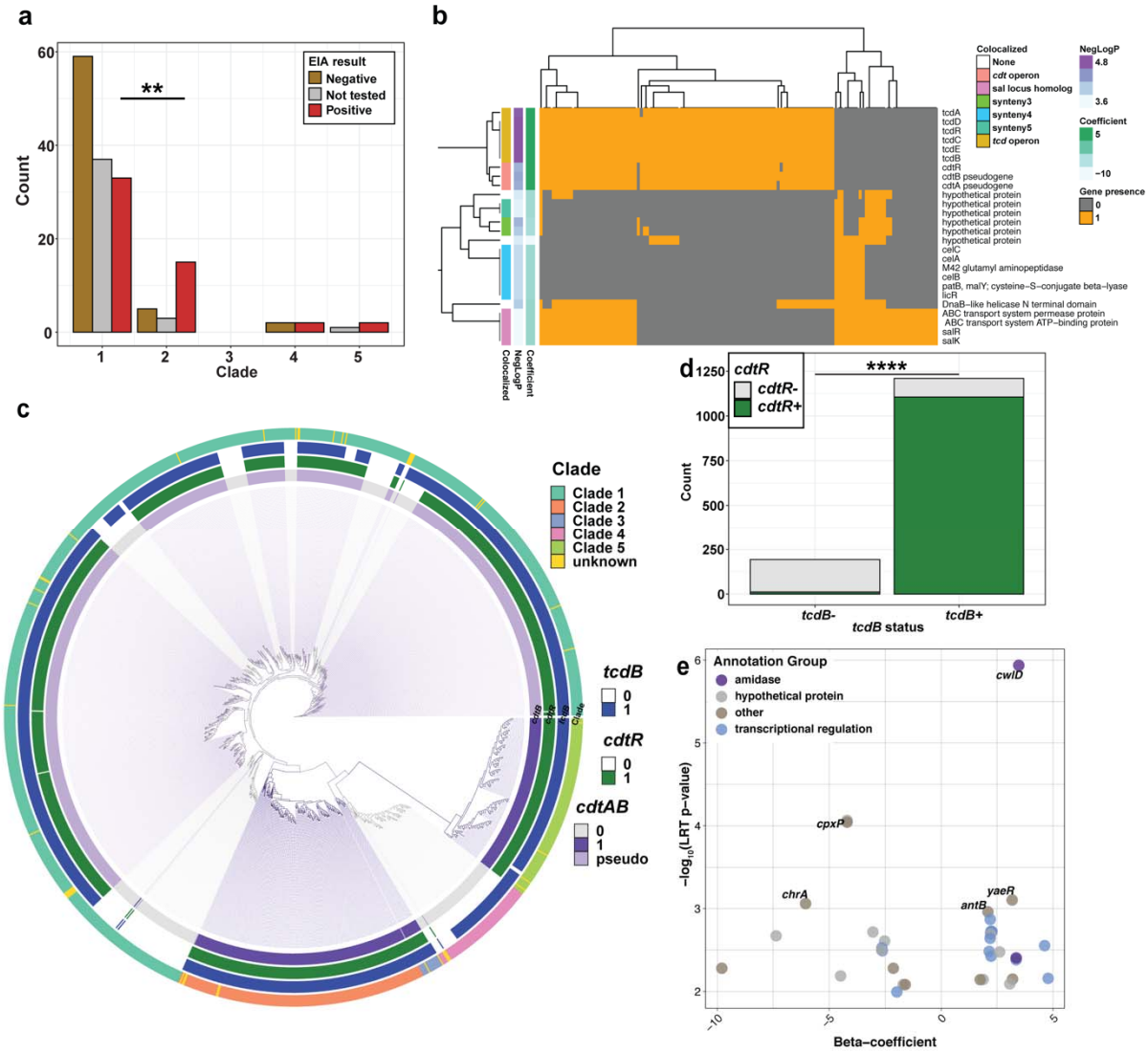
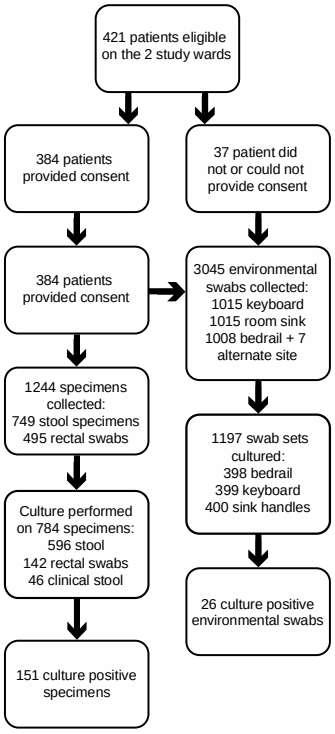


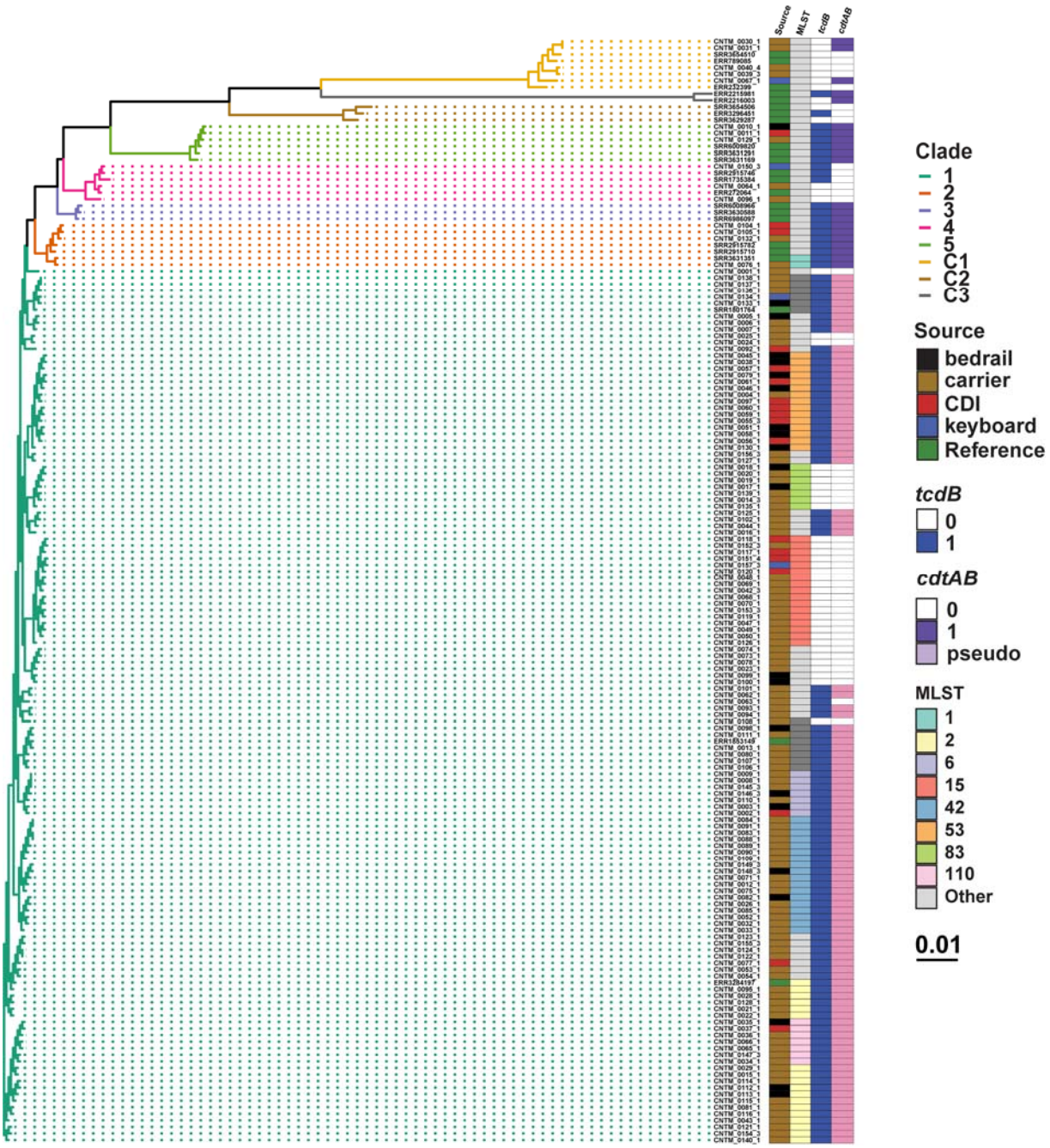
Figure 5: Clade 1 is responsible for the majority of CDI cases and carries unique correlates to symptom severity.

Caption: a) EIA status by clade across this and a previous study[29]. Fisher's exact test, p<0.01. c) Phylogenetic tree of >1400 *C. difficile* isolates from NCBI (Supplementary Table 4) depicting presence of binary toxin and PaLoc operons. d) Presence of full-length *cdtR* and association with *tcdB* presence. e) Filtered results (p-values <0.01) pyseer analysis evaluating gene association with CDI suspicion in Clade 1 isolates using the phylogenetically-corrected p-values (LRT). Purple color indicates p<0.001. Positive beta coefficient indicates gene association with CDI suspicion, while negative beta indicates asymptomatic colonization.

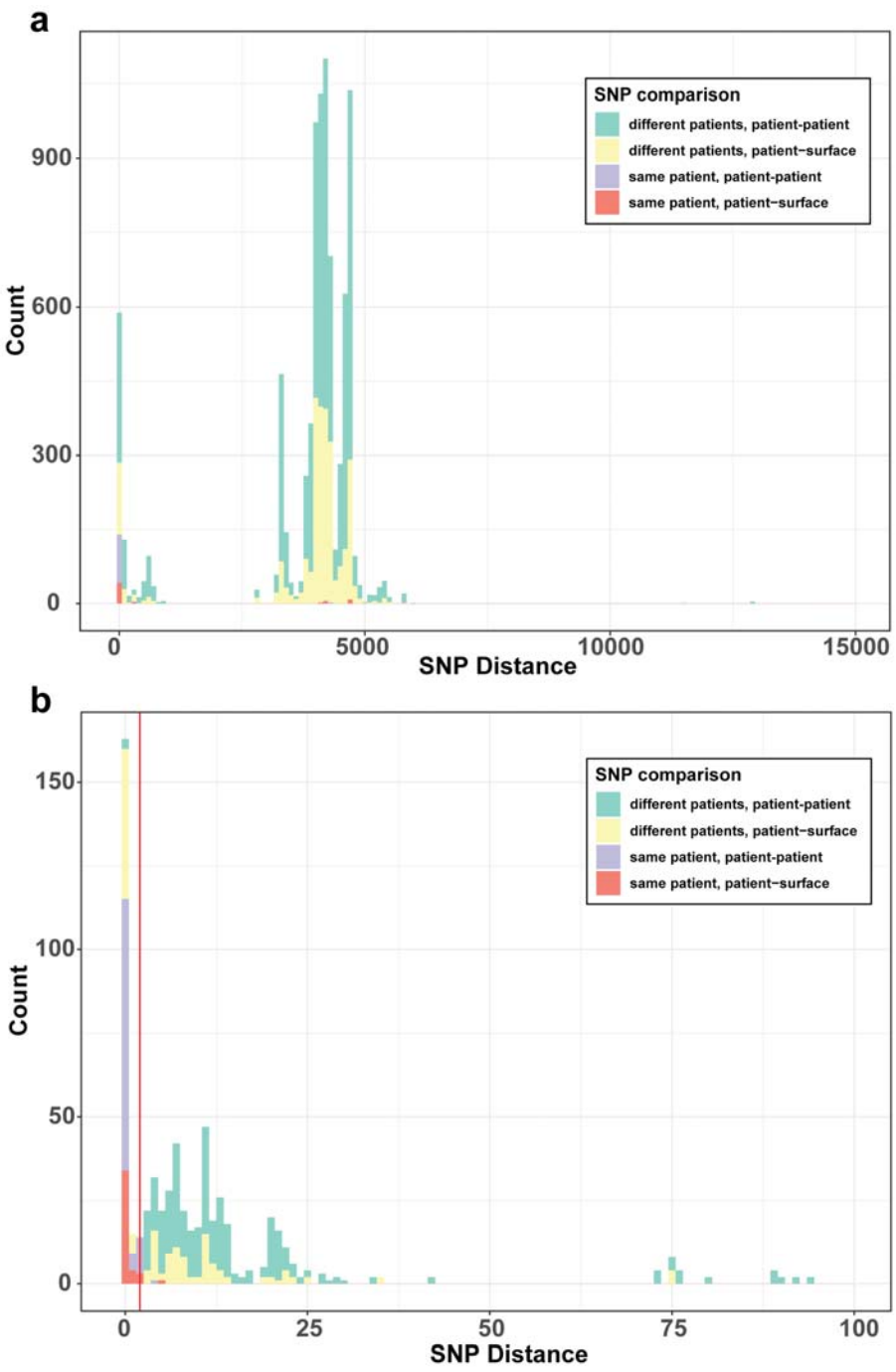
Supplemental Figures and Tables Titles and Captions



Supplementary Figure 1: Bubble plot of enrollment, collection, and culture numbers.



Supplementary Figure 2: Phylogenetic tree of isolates collected in this study and select references (Supplementary Table 2).



Supplementary Figure 3: Histogram of core genome SNP distances between different isolate comparisons, a) full histogram and b) zoomed to <200 SNPs.