Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

Can J Microbiol. Author manuscript; available in PMC 2026 May 15.

Published in final edited form as:

Can J Microbiol. 2026 January 01; 72: 1–11. doi:10.1139/cjm-2025-0288.

Pangenome-GWAS analysis of *Clostridioides difficile* resistant traits from healthcare and non-healthcare associated settings

Manuel Jara[1,*], **Alba Frias-De-Diego**[2], **Robert A. Petit III**[3], **Loc Nguyen**[4], **Cristina Lanzas**[5]

[1]School of Environmental Studies, College of Emerging and Integrative Studies, Tennessee Tech University, Cookeville, Tennessee, USA

[2]Department of Biology, College of Art and Sciences Tennessee Technological University, Cookeville, Tennessee, USA

[3]Division of Infectious Diseases, Department of Medicine, Emory University, Atlanta, Georgia, USA

[4]Department of Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg, Virginia, USA

[5]Department of Population Health and Pathobiology, College of Veterinary Medicine, North Carolina State University, Raleigh, NC, USA

Abstract

Clostridioides difficile (*C. difficile*) is a gastrointestinal pathogen that causes healthcare-associated infections with numerous mechanisms to acquire and confer antimicrobial resistance (AMR) and rapidly adapt to novel environments. These environments extend beyond healthcare, and encompass non-healthcare related settings, where antimicrobials engender distinct selection pressures. Employing a dataset of 3,870 whole-genome sequences of *C. difficile* from healthcare and non-healthcare environments, this study investigates the selection pressures associated with each setting and its impact on the prevalence of AMR genes of *C. difficile.* To achieve this, we used pangenome-wide association studies (pan-GWAS) and Generalized Linear Mixed Model to identify significant associations between AMR genes and the specific setting from which the isolates were collected. Our findings evidenced significant differences in the AMR gene repertoire between healthcare and non-healthcare associated *C. difficile* with antibiotics such as macrolides and fluoroquinolones being more linked with healthcare. Antibiotics commonly used in agriculture such as aminoglycosides or tetracyclines also presented associations with non-healthcare environments. Our work offers a comprehensive exploration of *C. difficile* AMR determinants across healthcare and non-healthcare related settings shedding light on the selection pressures present in each environment, enhancing our understanding of *C. difficile* resistome and its influence on the bacterium's opportunistic colonization nature.

[*]**Correspondence:** mjara@tntech.edu.

Conflict of interest statement

None to declare. The authors report no conflict of interest.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Keywords

C. difficile ; antimicrobial resistance genes; whole-genome sequencing; pangenome; pangwas; genomic epidemiology

Introduction

Clostridioides difficile (*C. difficile*) is an anaerobic, spore-forming gastrointestinal pathogen and one of the leading causes of healthcare-associated infections worldwide (He, Miyajima et al. 2013). *C. difficile* infection (CDI) is associated with a wide range of clinical manifestations such as diarrhea, toxic megacolon, and colon perforation in addition to asymptomatic carriage (Dubberke, Haslam et al. 2011). Antibiotics disrupt the colon microbiota, facilitating *C. difficile* growth and subsequent infection (Britton and Young 2014, Zhang, Dong et al. 2015). Recent antibiotic exposure is present in the vast majority of hospitalized patients developing CDI (Schechner, Fallach et al. 2021). Additionally, patients colonized and infected with *C. difficile* can contaminate their surrounding environment with spores and potentially expose healthcare workers to the pathogen. Spores can persist for a long time in the built environment and be transmitted to susceptible patients (Weber, Anderson et al. 2013). *C. difficile* has also emerged as a source of non-healthcare related infections. From an estimated total of 462,000 cases in 2017 in the United States, half of them were non-healthcare related with no documentation of prior admission to a healthcare facility in the preceding 12 weeks (Rupnik, Wilcox et al. 2009, Leffler and Lamont 2015, CDC 2019, Guh, Mu et al. 2020). In addition to colonizing humans, *C. difficile* has also been isolated from numerous animal hosts, including horses, pigs, cattle, dogs, cats, hamsters, guinea pigs, rats, and rabbits, as well as environmental reservoirs (Roy Chowdhury, DeMaere et al. 2016, Stone, Sidak-Loftis et al. 2016, Knight, Squire et al. 2017, Martín-Burriel, Andrés-Lasheras et al. 2017), suggesting additional suitable reservoirs of *C. difficile* besides the healthcare setting. Genomic studies indicate that certain *C. difficile* strains spread differently across healthcare and non-healthcare related settings, undergoing further adaptation to specific niches and environments (Knetsch, Connor et al. 2014, Eyre, Davies et al. 2018, Miles-Jay, Young et al. 2021).

C. difficile evolution has been characterized by the rapid acquisition of AMR. Resistance to commonly encountered antimicrobials may increase the time window in which *C. difficile* can colonize the gut (Baines and Wilcox 2015). For instance, the swift rise of the epidemic strain 027/B1/NAP1 in the early 2000s is attributed to its acquisition of fluoroquinolone resistance (He, Miyajima et al. 2013). Overall, *C. difficile* showcases an impressive ability to adapt swiftly and develop resistance across various antimicrobial classes, spanning aminoglycosides, β-lactams, fluoroquinolones, lincosamides, macrolides, and tetracyclines (Peng, Jin et al. 2017, Imwattana, Rodríguez et al. 2021, O'Grady, Knight et al. 2021). The mechanisms underpinning *C. difficile's* resistance are multifaceted, encompassing chromosomal resistance genes, alterations in metabolic pathways, mobile genetic elements (MGEs), point mutations, and biofilm formation (Spigaglia, Mastrantonio et al. 2018, O'Grady, Knight et al. 2021).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Comparative pangenome analysis is a novel approach to identifying genetic determinants that might be differentially associated with clinical disease and selective pressures (Hyun, Monk et al. 2022, Dey, Gaur et al. 2023). The pangenome represents the combination of all the genes existing within a specific species or a group of microorganisms and comprises both a central, or core genome, and an accessory genome (Medini, Donati et al. 2005, Croll and McDonald 2012).The core genome includes genes present in all strains, whereas the accessory genome encompasses genes that are particular to specific strains or even unique (Laing, Buchanan et al. 2010). Previous literature has highlighted the importance of the accessory genome due to its significant influence on phenotypic diversity and the evolution of the genome, including genes related to drug resistance (Zakham, Sironen et al. 2021). While many accessory loci are not essential under baseline conditions, they can underpin adaptation to specific ecological niches and selective pressures, including antimicrobial exposure (Vernikos, Medini et al. 2015, Lv, Bi et al. 2025). Operationally, accessory gene content can be described along a frequency gradient (e.g., soft-core, shell, and cloud), which helps distinguish near-universal from rare genes within a population (Matthews, Watson-Haigh et al. 2024).

For that reason, a pangenome genome-wide association study (pan-GWAS) is a powerful approach to studying selection pressures in medically important bacteria such as *C. difficile,* since it can reveal the genetic and functional diversity of bacterial populations and identify the genes and variants associated with phenotypic traits of interest (Fatoba, Fatoba et al. 2022, Golchha, Nighojkar et al. 2022, Hyun, Monk et al. 2022, Norsigian, Danhof et al. 2022, Zhou, Jiang et al. 2023). It can compare the genomes of thousands of isolates from different species, regions, hosts, or conditions, and detect the genes that are enriched or depleted in certain groups. It can also account for the phylogenetic relationships among the isolates and control for confounding factors such as population structure, recombination, and linkage disequilibrium (Liao, Guo et al. 2021, Harling-Lee, Gorzynski et al. 2022). This approach allows us to uncover the evolutionary forces that shape bacterial pangenomes, including horizontal gene transfer, gene loss, and selection. For those reasons, pan-GWAS serves as a robust and versatile tool for exploring the genetic basis of bacterial adaptation and resistance.

This study aims to assess the presence of potential selection pressures associated with healthcare and non-healthcare related settings by identifying what AMR genes are differentially enriched in those settings. To achieve this, we compiled an extensive dataset of whole-genome sequencing isolates, along with their associated metadata. We performed a pangenome genome-wide association study (pan-GWAS) to identify significant associations between AMR genes and the specific setting from which the isolates were collected (i.e., healthcare or non-healthcare related).

Materials and Methods

Figure 1 presents the complete analytical pipeline and includes the extraction and curation of the sequence data, pangenome analysis, and its post-process to identify associations between genomic features and the origin of the sequences. We focused on five prevalent sequence types (ST1, ST2, ST3, ST11, and ST15) because they provided sufficient sample

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

sizes and, critically, each included isolates from both Healthcare and External settings with interpretable isolation-source metadata, enabling within-ST comparisons. A total of 3,870 isolates of *C. difficile* were retrieved between 2000 and 2021 from the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) (Katz, Shutov et al. 2022) repository (see Table S1 for the complete list) with fastq-dl v2.0.4 (Petit III and Read 2020).

The raw reads were processed using the Bactopia v3 pipeline (Petit III and Read 2020), which involved sequence quality control and the removal of low-quality bases and reads (Q < 20) (Andrews 2010, Bushnell 2020, Petit III 2022). Metadata associated with each isolate included collection date, country of origin, isolation source, and NCBI bioproject (Table S1) from SRA/BioSample and EnteroBase and used it to classify sampling context. Collection date granularity varied across records and, for many isolates, only the calendar year was available. Isolates with missing or ambiguous isolation source metadata were excluded from source-stratified analyses, and synonymous source labels were standardized for consistency (e.g., stool/feces). For clarity, we use Healthcare for isolates collected within human clinical care settings (acute-care hospitals, outpatient clinics, emergency departments, rehabilitation units, and long-term care/nursing facilities), their built environments (e.g., patient rooms, equipment, sinks/drains), and clinical diagnostic laboratories processing human specimens; wastewater or drains sampled from healthcare facilities were also treated as healthcare. External (non-healthcare) refers to sampling outside healthcare facilities, including community human carriage sampled outside clinics, veterinary and animal sources (livestock, companion animals, wildlife), food and retail meat, farm/abattoir settings, and environmental matrices such as soil, sediments, surface waters, and municipal wastewater not explicitly linked to a healthcare facility, as well as fomites in public spaces. High-quality reads were then assembled using Shovill (v1.1.0), a tool integrated into the Bactopia pipeline that uses SPAdes v3.15.5 (Bankevich, Nurk et al. 2012, Seemann 2017). The quality of assemblies was assessed with Mash 2.3.0 and QUAST v5.2.0 (Gurevich, Saveliev et al. 2013, Ondov, Starrett et al. 2019). The Bactopia pipeline was then used to identify Single Nucleotide Polymorphisms (SNPs) between the isolates and a selected reference genome (AM180355.1), with Snippy v4.6.0 (https://github.com/tseemann/snippy), creating an alignment based on the variant sites.

Using the genomes annotated with Prokka v1.14.0 (Seemann 2014), we performed a pangenome analysis at the ST level, as determined by Ariba (v2.14.6) and BLAST+ (v2.16.0) (Camacho, Coulouris et al. 2009, Hunt, Mather et al. 2017) using Panaroo (v1.2.2) on clean-mode setting (Tonkin-Hill, MacAlasdair et al. 2020). Gene clusters were categorized by their frequency across genomes within each ST to distinguish highly conserved loci from those that are variable across the population. We defined core genes as gene clusters present in ≥99% of genomes within a given ST, soft-core as 95% to <99%, shell as 15% to <95%, and cloud as <15%.In this study, we considered STs 1, 2, 3, 11, and 15, and we also summarized pangenome composition for all the sequences combined (Table S1).

An initial maximum-likelihood phylogeny was inferred from the core-genome SNP alignment using IQ-TREE 2 (Minh, Schmidt et al. 2020). We then applied ClonalFrameML

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

(v1.12) (Didelot and Wilson 2015) to infer and account for homologous recombination and to reconstruct a recombination-corrected phylogeny; this tree was used in downstream pan-GWAS. All the annotated sequences were also scanned for AMR determinants using AMRFinderPlus (v4.0.3) (Feldgarden, Brover et al. 2019), and we retained only hits with ≥99% identity and ≥99% coverage to minimize false-positive calls.

To determine whether AMR determinants differed between sampling contexts (i.e., healthcare and non-healthcare related) at antibiotic class level, we used two different approaches: i) Generalized Linear Mixed Model (GLMM) to assess the association between antimicrobial resistance (AMR) genes and the source environment (healthcare vs external). For GLMMs, the presence or absence of each AMR determinants (binary response variable) was modeled as a function of the environmental source (fixed effect), with the year of data collection and country of origin included as random factors to account for temporal and geographic heterogeneity. Because collection dates were frequently only available at the calendar-year level and sampling was uneven through time, we did not model ordered temporal dependence or seasonality. By conducting separate analyses for each ST, we aimed to control for potential differences in sampling and evolutionary history between the STs. The GLMM used a binomial distribution with a logit link function, and models were fit using the lme4 package in R (Bates, Mächler et al. 2014) and ii) a pangenome genome-wide association study (pan-GWAS). We determine as healthcare-related samples all isolates collected within a healthcare-associated facility such as hospitals, clinics, and nursing homes. We included within the non-healthcare associated isolates any samples coming from livestock production, agriculture, wildlife, pets, and any other origin outside of a healthcare facility. This approach expands the scope of GWAS analyses by considering genetic variations present in the pangenome, which allows to capture a broader range of genetic diversity within species while accounting for population structure (Agarwal, Choudhary et al. 2021, Gupta 2021). To do this, Scoary implements a pairwise comparison algorithm (Maddison 2000) which requires a phylogenetic tree.

In our study, pan-GWAS analyses were conducted using the Scoary2 (v0.0.15) approach (Roder, Pimentel et al. 2023), which is a robust method implemented from a previous version “Scoary” (Brynildsrud, Bohlin et al. 2016). Scoary2 uses Fisher’s test to determine if there are associations between two categorical variables by combining information from gene presence or absence patterns (obtained through the Panaroo pangenome analysis) with trait data (the setting from which each isolate originated) to identify genomic elements associated with specific phenotypic variations. Due to the large number of genes under examination, Scoary2 corrected the *p*-values from Fisher’s test for multiple testing, yielding *q*-values (Storey 2003). This analysis enables the detection of genes that may be responsible for the observed differences in traits among the strains. By integrating genotypic and phenotypic data, we aimed to uncover significant associations at the gene level taking into account population structure (using a phylogenetic tree, which accounted for recombination, that was obtained from ClonalFrameML) to avoid the above-mentioned bias. The strength of this association was quantified through odds ratio values. To ensure that our analyses are robust and accurate, we performed the pangenome and pan-GWAS analyses at the sequence type (ST) level. This decision was made because each sequence type has been sampled differently mostly based on its pathogenicity and evolutionary history.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Results

Of the 3,870 *C. difficile* isolates analyzed, 85.8% of the sequences came from healthcare settings, while 14.2% were associated with non-healthcare settings (such as agriculture, livestock, pets, and other non-healthcare origins) (Figure 2A). The samples come from 34 countries, with the majority (75%) coming from the United States, Germany, and Australia (Figure 2B–C). An examination of the temporal distribution revealed that healthcare-related samples presented a sharp increase starting around 2005, peaking between 2015 and 2017, followed by a decline after 2017, while the non-healthcare samples show a flatter trend over time with occasional peaks around 2010 and 2015 but remain significantly lower in number compared to healthcare samples (Figure 2D). The isolates were classified into five distinct sequence types (STs), ST1 (31.0%), ST2 (30.6%), ST3 (12.4%), ST11 (21.3%), and ST15 (4.7%) (Figure 2E) (Table S1).

C. difficile's resistance gene variations

The *C. difficile* resistome comprised a total of 112 AMR determinants (acquired genes plus defined point mutations/allelic variants) conferring resistance to 21 antibiotic classes. Among the genes identified, the most representative classes were β-lactams, where the *blaCDD* gene was the most frequently detected, with a prevalence of 95.1%. Additionally, the *mreE* (V497L) allelic variant was present in 31.0% of the samples. Quinolone-associated mutations were also common, including *gyrA* (T82I) and *gyrB* (I139R) in 40.2% and 30.6% of the samples respectively. In addition, MLS (Macrolide-Lincosamide-Streptogramin) resistance was observed in 22.8% of the samples due to the presence of the *erm*(B) gene (Table S1).

At the sequence type (ST) level, the GLMM and pan-GWAS analyses revealed relevant variations in AMR determinant patterns associated with the two different settings (Figures 3–7). For example, ST1 presented high versatility, displaying resistance against 12 AMR classes. For β-lactam resistance, the *mreE* (A778V) gene was significantly more prevalent in isolates from non-healthcare settings, similar to what we found in the *vanR-Cd* (T115A) determinant was the only one supported by both GLMM ($p<0.01$) and pan-GWAS analyses ($q^*ep < 0.001$, odds ratio: 3.7). In contrast, determinants more strongly associated with healthcare settings included *erm(B)* (macrolide-lincosamide-streptogramin) and gyrA (T82I) associated with quinolone resistance (Figure 3).

In the case of ST2, the most prevalent determinants found in both settings were *blaCDD* (β-lactams) and *gyrB* (I139R) (Quinolones) which were found in over 95% of the isolates. However, when comparing the determinant prevalence between the two different settings, we found that all the determinant that showed significant differences were more prevalent in non-healthcare settings. Some of those genes were represented by several point mutations on the *mreE* gene (β-lactam resistance), as well as the presence of the *tet*(M), and the *dfrF* genes, which confers resistance to tetracyclines and trimethoprim respectively. In this case, only two determinants (*mreE(A694V),* and *tet(M)* which confer resistance to β-lactams and tetracyclines respectively, presented significant results for both GLMM and pan-GWAS analyses ($p<0.001$; odds ratio 3.3 and 7.1 respectively) (Figure 4).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

For ST3, the β-lactam gene *blaCDD* was the most prevalent across the samples, present in all of them. Contrary to what was observed in ST2, most of the differences in gene presence between settings showed a higher association with healthcare. This was represented by the genes *mreE* (V497L) and *erm*(B), which confer resistance against β-lactams and MLS respectively. In this ST, only the *erm(B)* gene which confers resistance against MLS presented significant differences in both GLMM and pan-GWAS analyses ($p<0.01$; odds ratio = 18.1) (Figure 5).

For ST11, we identified resistance to 19 different antimicrobial classes, with aminoglycosides, quinolones, glycopeptides, and tetracyclines showing the highest number of genes and point mutations. At the gene level, *blaCDD, gyrB* (S366V), and *gyrB* (S416A) were the most prevalent genes (>99%). When comparing the proportion of AMR determinant prevalence between healthcare and non-healthcare settings, we found significant differences, with most resistance genes showing higher prevalence in healthcare. For example, the genes *aad9* and *spw*, which confer resistance against aminoglycosides, the *erm*(B) gene (MLS resistance), the *gyrA* (T82I) (quinolone resistance), and the *tet*(M) gene (tetracyclines resistance). In contrast, the aminoglycoside resistance gene *aadE* was more prevalent in non-healthcare settings. In this ST, we found three determinants that presented significant results for both the GLMM and pan-GWAS analyses: *aad9* (aminoglycosides; $p<0.001$; odds ratio = 8.0), *erm(B)* (MLS; $p<0.001$; odds ratio 28.1), and *tet(M)* (tetracyclines; $p<0.001$; odds ratio = 2.8) (Figure 6).

Moving on to ST15, we observed determinants conferring resistance against 10 different resistance classes. The most prevalent gene found was *blaCDD-2* (present in over 80% of the isolates). Several genes evidenced significant differences associated with the origin of the samples, for example, *aadE* gene (aminoglycoside resistance), *erm(B)* (MLS resistance), and *tet(M)* (tetracycline resistance) showed higher prevalence in healthcare settings. On the other hand, *blaCDD-2* and *mreE* (A795V) (β-lactam resistance) (showed a higher presence in non-healthcare isolates. In this case, only two determinants (*aadE;* $p<0.01$; odds ratio=3.2; and *blaCDD-2*; $p<0.01$; odds ratio 33.7), conferring resistance to aminoglycosides and β-lactams respectively presented significant results for both the GLMM and pan-GWAS analyses (Figure 7).

Discussion

This study advances our understanding of *C. difficile* AMR dynamics in healthcare and non-healthcare environments and highlights the evolutionary adaptability of this bacterium to different selective pressures. The utilization of pangenome-wide association studies (pan-GWAS) expanded on prior research by allowing the comparison of resistance profiles between isolates from healthcare and non-healthcare isolates. Previous studies were limited to the study of clinical isolates (Imwattana, Rodríguez et al. 2021, Miles-Jay, Young et al. 2021), however, our study was able to identify rare resistance genes that may be associated with environmental-specific adaptations. By detecting the differences in AMR gene profile between healthcare and non-healthcare settings, this study adds context to the *C. difficile* resistome, suggesting specific genetic adaptations shaping bacterial persistence and spread within and beyond healthcare settings.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Identifying specific associations between AMR determinants and the environment could inform future investigations on transmission dynamics and control measures for *C. difficile*, particularly in settings where antibiotic use can drastically vary. The AMR patterns found in this study provide a foundation for intervention that considers specific AMR determinants. For instance, the observed notable predominance of sequences originating from healthcare settings and distinct selection pressures favored resistance to antibiotics commonly used in clinical practice such as β-lactams, macrolides, and aminoglycosides. However, these antibiotics are not typically used for treating CDI but are commonly administered in healthcare settings for other bacterial infections (Surawicz, Brandt et al. 2013). The potential fitness benefits of the acquisition of these particular AMR determinants likely lies in their ability to enhance *C. difficile* colonization in the gut. Broad-spectrum antibiotics disrupt the gut microbiome creating a favorable environment for the colonization of *C. difficile* (Schaeffler and Breitrueck 2018). Resistance to these antibiotics may provide an additional competitive advantage for *C. difficile* colonization (Baines and Wilcox 2015, O'Grady, Knight et al. 2021). The primary antibiotics currently used to treat CDI are metronidazole and vancomycin (Griffin and Arnold 2011, Katzman 2014), followed by fidaxomicin which is a promising option to treat patients at high risk for CDI relapse (Shehabi, Badran et al. 2015). These antibiotics, however, did not exhibit significant associations with healthcare environments, which suggests that the main antibiotic-related selective pressure for *C. difficile* evolution comes as an indirect effect of antibiotic use for other bacterial species.

Resistance against other antimicrobial classes was associated with non-healthcare settings, which can be associated with the different use of antimicrobials for activities such as agriculture or veterinary medicine. For example, tetracyclines are extensively used in livestock, agriculture, and aquaculture, surpassing their application in human healthcare (Eliopoulos, Eliopoulos et al. 2003, Roberts 2005, Roberts and Schwarz 2016). Most of these determinants have been described to spread within and between bacterial species via horizontal gene transfer (Von Wintersdorff, Penders et al. 2016, Evans, Griffith et al. 2020, Michaelis and Grohmann 2023). Therefore, the application of whole-genome approaches, and particularly pan-GWAS, is critical to identify mobile genetic elements that facilitate horizontal gene transfer and thus, better understand *C. difficile* AMR dynamics.

The consideration of non healthcare-associated data is critical for the holistic understanding of *C. difficile* AMR dynamics and how this pathogen can adapt to varied selection pressures, highlighting its ability to survive and spread in non-healthcare environments such as agricultural settings, which are important sources of AMR determinants diversity (Jadeja and Worrich 2022, Kaviani Rad, Balasundram et al. 2022, Miller, Ferreira et al. 2022). The epidemiological insights of this study set a valuable precedent for future studies in other pathogens and their relationship with the environment in the context of AMR.

While this study provides significant insights into *C. difficile* epidemiology, it is crucial to acknowledge certain limitations. The inherent bias in sampling arising from the predominance of healthcare-associated isolates highlights the need for broader sampling strategies considering non-healthcare settings to ensure a more balanced representation of the different sampling sources. In addition, collection dates were frequently incomplete (often available only at the year level) and unevenly distributed through time; accordingly,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

year was modeled as a categorical random effect and we did not explicitly test for seasonal patterns, despite evidence that CDI incidence can exhibit seasonality in multiple settings (Furuya-Kanamori, McKenzie et al. 2015). Additionally, given the nature of the data, there was no available information to differentiate toxigenic and non-toxigenic isolates beforehand, as happened with the susceptibility of isolates independently of gene carriage. Moreover, investigating the functional impact of genetic variations and expression patterns in response to specific environments would enhance our understanding of the mechanisms driving adaptation and survival. Finally, it is important to consider that although our database is extensive, it may not be fully representative of the current *C. difficile* diversity due to its rapid evolution and adaptation, especially in non-healthcare environments.

Our study provides a comprehensive assessment of the AMR landscape in *C. difficile* in healthcare and non-healthcare environments, revealing specific AMR determinants-environment associations that highlight the bacterial adaptive capacity. Using a pan-GWAS approach, this study identified specific AMR patterns shaped by specific selective pressures found in healthcare and non-healthcare settings. As additional genomic data becomes available, this work will serve as a foundation for the integration of this information into research aimed at studying the ecological drivers of AMR evolution, bringing global health implications for combating AMR.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

This work was supported by the Centers for Disease Control and Prevention (CDC) (U01CK000587) and the US National Institutes of Health (NIH) (R35GM134934). Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funders.

Data sharing

The NCBI-accession data including the available metadata for all the *C. difficile* sequences are included in the supplementary material. Here are the NCBI-Bioprojects used in this project: PRJEB11868, PRJEB12715, PRJEB12970, PRJEB13264, PRJEB13566, PRJEB14038, PRJEB19597, PRJEB20216, PRJEB22563, PRJEB22784, PRJEB23450, PRJEB26512, PRJEB26604, PRJEB27277, PRJEB28391, PRJEB32765, PRJEB36956, PRJEB37809, PRJEB3957, PRJEB40974, PRJEB41588, PRJEB43401, PRJEB43517, PRJEB43620, PRJEB44406, PRJEB45248, PRJEB47018, PRJEB5003, PRJEB6286, PRJEB6566, PRJEB6575, PRJEB7326, PRJEB8702, PRJEB8741, PRJEB9067, PRJNA227075, PRJNA227078, PRJNA227079, PRJNA231221, PRJNA232267, PRJNA238844, PRJNA239262, PRJNA239264, PRJNA239265, PRJNA271048, PRJNA274772, PRJNA278292, PRJNA283119, PRJNA296517, PRJNA305247, PRJNA309189, PRJNA314038, PRJNA314712, PRJNA321988, PRJNA325243, PRJNA328573, PRJNA342261, PRJNA342265, PRJNA342267, PRJNA342268, PRJNA342273, PRJNA342275, PRJNA342333, PRJNA342401, PRJNA342405, PRJNA342579, PRJNA342580, PRJNA342584, PRJNA342585,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

PRJNA342587, PRJNA342631, PRJNA342644, PRJNA342652, PRJNA342758, PRJNA342759, PRJNA342760, PRJNA342763, PRJNA342766, PRJNA342768, PRJNA342769, PRJNA342771, PRJNA344770, PRJNA379066, PRJNA379525, PRJNA381384, PRJNA398796, PRJNA430032, PRJNA438482, PRJNA445766, PRJNA449640, PRJNA472399, PRJNA475751, PRJNA478136, PRJNA481169, PRJNA486358, PRJNA490071, PRJNA490820, PRJNA494608, PRJNA526195, PRJNA530817, PRJNA545620, PRJNA555597, PRJNA561087, PRJNA577141, PRJNA590803, PRJNA591265, PRJNA594943, PRJNA595724, PRJNA599117, PRJNA600974, PRJNA629351, PRJNA630469, PRJNA646752, PRJNA658263, PRJNA658847, PRJNA659471, PRJNA663277, PRJNA672122, PRJNA673132, PRJNA679767, PRJNA736231, PRJNA762329, PRJNA785426, PRJNA789534, PRJNA789842, PRJNA85689, PRJNA85693, PRJNA85697, PRJNA85713, PRJNA85729, PRJNA85731, PRJNA85735, PRJNA85763, PRJNA85783, PRJNA85787, PRJNA85795, PRJNA85799, PRJNA85801, PRJNA85811, PRJNA85819, PRJNA85823, PRJNA85829, PRJNA85881, PRJNA85893, PRJNA85895, PRJNA85897, PRJNA85909, PRJNA85959, PRJNA85989, PRJNA86005

References

Agarwal G, Choudhary D, Stice SP, Myers BK, Gitaitis RD, Venter SN, Kvitko BH and Dutta B (2021). “Pan-genome-wide analysis of Pantoea ananatis identified genes linked to pathogenicity in onion.” Frontiers in Microbiology: 2381. 10.3389/fmicb.2021.684756.

Andrews S (2010). FastQC: a quality control tool for high throughput sequence data, Babraham Bioinformatics, Babraham Institute, Cambridge, United Kingdom.

Baines SD and Wilcox MH (2015). “Antimicrobial resistance and reduced susceptibility in Clostridium difficile: potential consequences for induction, treatment, and recurrence of C. difficile infection.” Antibiotics 4(3): 267–298. 10.3390/antibiotics4030267. [PubMed: 27025625]

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, Lesin VM, Nikolenko SI, Pham S and Prjibelski AD (2012). “SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing.” Journal of computational biology 19(5): 455–477. 10.1089/cmb.2012.002. [PubMed: 22506599]

Bates D, Mächler M, Bolker B and Walker S (2014). “Fitting linear mixed-effects models using the lme4 package in R.” J Stat Softw 67: 1–48.

Britton RA and Young VB (2014). “Role of the intestinal microbiota in resistance to colonization by *Clostridium difficile*.” Gastroenterology 146(6): 1547–1553. 10.1053/j.gastro.2014.01.059. [PubMed: 24503131]

Brynildsrud O, Bohlin J, Scheffer L and Eldholm V (2016). “Rapid scoring of genes in microbial pan-genome-wide association studies with Scoary.” Genome biology 17: 1–9. [PubMed: 26753840]

Bushnell B (2020). “BB Map Short Read Aligner, and Other Bioinformatic Tools.” Available from: sourceforge.net/projects/bbmap. 10.1186/s13059-016-1108-8.

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K and Madden TL (2009). “BLAST+: architecture and applications.” BMC bioinformatics 10: 1–9. [PubMed: 19118496]

CDC (2019). “CDC’s antibiotic resistance threats in the United States, 2019.” Centers for Disease Control and Prevention, Atlanta, GA. 10.1186/1471-2105-10-421.

Corredor M, Patiño-Salazar JD, Castaño DC and Muñoz-Gómez A (2023). “The Pangenome of Pseudomonas aeruginosa.” In *Pseudomonas aeruginosa*-New Perspectives and Applications. IntechOpen. 10.5772/intechopen.108187.

Croll D and McDonald BA (2012). “The accessory genome as a cradle for adaptive evolution in pathogens.” PLoS pathogens 8(4): e1002608. 10.1371/journal.ppat.1002608. [PubMed: 22570606]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Dey S, Gaur M, Sykes EM, Prusty M, Elangovan S, Dixit S, Pati S, Kumar A and Subudhi E (2023). "Unravelling the Evolutionary Dynamics of High-Risk *Klebsiella pneumoniae* ST147 Clones: Insights from Comparative Pangenome Analysis." Genes 14(5): 1037. 10.3390/genes14051037. [PubMed: 37239397]

Didelot X and Wilson DJ (2015). "ClonalFrameML: efficient inference of recombination in whole bacterial genomes." PLoS computational biology 11(2): e1004041. 10.1371/journal.pcbi.1004041. [PubMed: 25675341]

Dubberke ER, Haslam DB, Lanzas C, Bobo LD, Burnham CA, Gröhn YT and Tarr PI (2011). "The ecology and pathobiology of *Clostridium difficile* infections: an interdisciplinary challenge." Zoonoses and public health 58(1): 4–20. 10.1111/j.1863-2378.2010.01352.x. [PubMed: 21223531]

Eliopoulos GM, Eliopoulos GM and Roberts MC (2003). "Tetracycline therapy: update." Clinical infectious diseases 36(4): 462–467. 10.1086/367622. [PubMed: 12567304]

Evans DR, Griffith MP, Sundermann AJ, Shutt KA, Saul MI, Mustapha MM, Marsh JW, Cooper VS, Harrison LH and Van Tyne D (2020). "Systematic detection of horizontal gene transfer across genera among multidrug-resistant bacteria in a single hospital." Elife 9: e53886. 10.7554/eLife.53886. [PubMed: 32285801]

Eyre DW, Davies KA, Davis G, Fawley WN, Dingle KE, De Maio N, Karas A, Crook DW, Peto TE and Walker AS (2018). "Two distinct patterns of *Clostridium difficile* diversity across Europe indicating contrasting routes of spread." Clinical Infectious Diseases 67(7): 1035–1044. 10.1093/cid/ciy252. [PubMed: 29659747]

Fatoba AJ, Fatoba DO and Babalola SO (2022). "Pangenome and subtractive genomic analysis of *Clostridioides difficile* reveals putative drug targets." Journal of Proteins and Proteomics 13(4): 247–256. 10.1007/s42485-022-00097-y.

Feldgarden M, Brover V, Haft DH, Prasad AB, Slotta DJ, Tolstoy I, Tyson GH, Zhao S, Hsu C-H and McDermott PF (2019). "Validating the AMRFinder tool and resistance gene database by using antimicrobial resistance genotype-phenotype correlations in a collection of isolates." Antimicrobial agents and chemotherapy 63(11): 10.1128/aac.00483-19.

Furuya-Kanamori L, McKenzie SJ, Yakob L, Clark J, Paterson DL, Riley TV and Clements AC (2015). "*Clostridium difficile* infection seasonality: patterns across hemispheres and continents–A systematic review." PloS one 10(3): e0120730. 10.1371/journal.pone.0120730. [PubMed: 25775463]

Golchha NC, Nighojkar A and Nighojkar S (2022). "Redefining genomic view of *Clostridioides difficile* through pangenome analysis and identification of drug targets from its core genome." Drug Target Insights 16: 17. 10.33393/dti.2022.2469. [PubMed: 36415217]

Griffin AT and Arnold FW (2011). "The most effective treatments for *Clostridium difficile* diarrhea: an evidence-based review." Infectious Diseases in Clinical Practice 19(3): 171–181. 10.1097/IPC.0b013e31820fc87d.

Guh AY, Mu Y, Winston LG, Johnston H, Olson D, Farley MM, Wilson LE, Holzbauer SM, Phipps EC and Dumyati GK (2020). "Trends in US burden of *Clostridioides difficile* infection and outcomes." New England Journal of Medicine 382(14): 1320–1330. 10.1056/NEJMoa1910215. [PubMed: 32242357]

Gupta PK (2021). "Quantitative genetics: pan-genomes, SVs, and k-mers for GWAS." Trends in Genetics 37(10): 868–871. 10.1016/j.tig.2021.05.006. [PubMed: 34183185]

Gurevich A, Saveliev V, Vyahhi N and Tesler G (2013). "QUAST: quality assessment tool for genome assemblies." Bioinformatics 29(8): 1072–1075. 10.1093/bioinformatics/btt086. [PubMed: 23422339]

Harling-Lee JD, Gorzynski J, Yebra G, Angus T, Fitzgerald JR and Freeman TC (2022). "A graph-based approach for the visualisation and analysis of bacterial pangenomes." BMC bioinformatics 23(1): 1–15. 10.1186/s12859-022-04898-2. [PubMed: 34983366]

He M, Miyajima F, Roberts P, Ellison L, Pickard DJ, Martin MJ, Connor TR, Harris SR, Fairley D and Bamford KB (2013). "Emergence and global spread of epidemic healthcare-associated *Clostridium difficile*." Nature genetics 45(1): 109–113. 10.1038/ng.2478. [PubMed: 23222960]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Hunt M, Mather AE, Sánchez-Busó L, Page AJ, Parkhill J, Keane JA and Harris SR (2017). "ARIBA: rapid antimicrobial resistance genotyping directly from sequencing reads." Microbial genomics 3(10). 10.1099/mgen.0.000131.

Hyun JC, Monk JM and Palsson BO (2022). "Comparative pangenomics: analysis of 12 microbial pathogen pangenomes reveals conserved global structures of genetic and functional diversity." BMC genomics 23(1): 1–18. 10.1186/s12864-021-08223-8. [PubMed: 34979896]

Imwattana K, Rodríguez C, Riley TV and Knight DR (2021). "A species-wide genetic atlas of antimicrobial resistance in *Clostridioides difficile*." Microbial Genomics 7(11). 10.1099/mgen.0.000696.

Jadeja NB and Worrich A (2022). "From gut to mud: dissemination of antimicrobial resistance between animal and agricultural niches." Environmental Microbiology 24(8): 3290–3306. 10.1111/1462-2920.15927. [PubMed: 35172395]

Katz K, Shutov O, Lapoint R, Kimelman M, Brister JR and O'Sullivan C (2022). "The sequence read archive: a decade more of explosive growth." Nucleic acids research 50(D1): D387–D390. 10.1093/nar/gkab1053. [PubMed: 34850094]

Katzman M (2014). "Antibiotic therapy for *Clostridium difficile* infection." Seminars in Colon and Rectal Surgery. 25(3): 143–149. 10.1053/j.scrs.2014.05.009.

Kaviani Rad A, Balasundram SK, Azizi S, Afsharyzad Y, Zarei M, Etesami H and Shamshiri RR (2022). "An overview of antibiotic resistance and abiotic stresses affecting antimicrobial resistance in agricultural soils." International Journal of Environmental Research and Public Health. 19(8): 4666. 10.3390/ijerph19084666. [PubMed: 35457533]

Knetsch C, Connor TR, Mutreja A, Van Dorp S, Sanders I, Browne H, Harris D, Lipman L, Keessen E and Corver J (2014). "Whole genome sequencing reveals potential spread of *Clostridium difficile* between humans and farm animals in the Netherlands, 2002 to 2011." Eurosurveillance. 19(45): 20954. 10.2807/1560-7917.es2014.19.45.20954. [PubMed: 25411691]

Knight DR, Squire MM, Collins DA and Riley TV (2017). "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." Frontiers in microbiology. 7: 2138. 10.3389/fmicb.2016.02138. [PubMed: 28123380]

Laing C, Buchanan C, Taboada EN, Zhang Y, Kropinski A, Villegas A, Thomas JE and Gannon VP (2010). "Pan-genome sequence analysis using Panseq: an online tool for the rapid analysis of core and accessory genomic regions." BMC bioinformatics. 11(1): 1–14. 10.1186/1471-2105-11-461 [PubMed: 20043860]

Leffler DA and Lamont JT (2015). "*Clostridium difficile* infection." New England Journal of Medicine. 372(16): 1539–1548. [PubMed: 25875259]

Liao J, Guo X, Weller DL, Pollak S, Buckley DH, Wiedmann M and Cordero OX (2021). "Nationwide genomic atlas of soil-dwelling Listeria reveals effects of selection and population ecology on pangenome evolution." Nature Microbiology. 6(8): 1021–1030. 10.1038/s41564-021-00935-7.

Lv T, Bi X, Zheng L, Zhao Y, Zhou Y, Wu T, Shen P, Zhu D, Chen S and Chen Y (2025). "Mobile genetic elements mediating antimicrobial resistance drive the evolutionary process of *Clostridioides difficile* ST37/RT017." BMC genomics. 26(1): 659. 10.1186/s12864-025-11822-4. [PubMed: 40652164]

Maddison WP (2000). "Testing character correlation using pairwise comparisons on a phylogeny." Journal of Theoretical Biology. 202(3): 195–204. 10.1006/jtbi.1999.1050. [PubMed: 10660474]

Martín-Burriel I, Andrés-Lasheras S, Harders F, Mainar-Jaime R, Ranera B, Zaragoza P, Falceto V, Bolea Y, Kuijper E and Bolea R (2017). "Molecular analysis of three *Clostridium difficile* strain genomes isolated from pig farm-related samples." Anaerobe. 48: 224–231. 10.1016/j.anaerobe.2017.09.002. [PubMed: 28928035]

Matthews CA, Watson-Haigh NS, Burton RA and Sheppard AE (2024). "A gentle introduction to pangenomics." Briefings in Bioinformatics 25(6): bbae588. 10.1093/bib/bbae588.

Medini D, Donati C, Tettelin H, Masignani V and Rappuoli R (2005). "The microbial pan-genome." Current opinion in genetics & development. 15(6): 589–594. 10.1016/j.gde.2005.09.006. [PubMed: 16185861]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Michaelis C and Grohmann E (2023). “Horizontal gene transfer of antibiotic resistance genes in biofilms.” Antibiotics. 12(2): 328. 10.3390/antibiotics12020328. [PubMed: 36830238]

Miles-Jay A, Young VB, Pamer EG, Savidge TC, Kamboj M, Garey KW and Snitkin ES (2021). “A multisite genomic epidemiology study of *Clostridioides difficile* infections in the USA supports differential roles of healthcare versus community spread for two common strains.” Microbial Genomics. 7(6). 10.1099/mgen.0.000590.

Miller SA, Ferreira JP and LeJeune JT (2022). “Antimicrobial use and resistance in plant agriculture: a one health perspective.” Agriculture. 12(2): 289. 10.3390/agriculture12020289.

Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A and Lanfear R (2020). “IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era.” Molecular biology and evolution. 37(5): 1530–1534. 10.1093/molbev/msaa015. [PubMed: 32011700]

Norsigian CJ, Danhof HA, Brand CK, Midani FS, Broddrick JT, Savidge TC, Britton RA, Palsson BO, Spinler JK and Monk JM (2022). “Systems biology approach to functionally assess the *Clostridioides difficile* pangenome reveals genetic diversity with discriminatory power.” Proceedings of the National Academy of Sciences. 119(18): e2119396119. 10.1073/pnas.2119396119.

O’Grady K, Knight DR and Riley TV (2021). “Antimicrobial resistance in *Clostridioides difficile*.” European Journal of Clinical Microbiology & Infectious Diseases. 40(12): 2459–2478. [PubMed: 34427801]

Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB and Phillippy AM (2019). “Mash Screen: high-throughput sequence containment estimation for genome discovery.” Genome biology. 20: 1–13. 10.1186/s13059-019-1841-x. [PubMed: 30606230]

Peng Z, Jin D, Kim HB, Stratton CW, Wu B, Tang Y-W and Sun X (2017). “Update on antimicrobial resistance in *Clostridium difficile*: resistance mechanisms and antimicrobial susceptibility testing.” Journal of clinical microbiology. 55(7): 1998–2008. 10.1128/jcm.02250-16. [PubMed: 28404671]

Petit RA III (2022). Fastq-scan: generate summary statistics of input FASTQ sequences. Github.

Petit RA III and Read TD (2020). “Bactopia: a flexible pipeline for complete analysis of bacterial genomes.” Msystems. 5(4): 10.1128/msystems.00190-00120. 10.1128/msystems.00190-20.

Roberts MC (2005). “Update on acquired tetracycline resistance genes.” FEMS microbiology letters. 245(2): 195–203. 10.1016/j.femsle.2005.02.034. [PubMed: 15837373]

Roberts MC and Schwarz S (2016). “Tetracycline and phenicol resistance genes and mechanisms: importance for agriculture, the environment, and humans.” Journal of environmental quality. 45(2): 576–592. 10.2134/jeq2015.04.0207. [PubMed: 27065405]

Roder T, Pimentel G, Fuchsmann P, Tena Stern M, von Ah U, Vergeres G, Peischl S, Brynildsrud O, Bruggmann R and Baur C (2023). “Scoary2: Rapid association of phenotypic multi-omics data with microbial pan-genomes.” Genome Biology. 25(93). 10.1186/s13059-024-03233-7.

Roy Chowdhury P, DeMaere M, Chapman T, Worden P, Charles IG, Darling AE and Djordjevic SP (2016). “Comparative genomic analysis of toxin-negative strains of *Clostridium difficile* from humans and animals with symptoms of gastrointestinal disease.” BMC microbiology. 16(1): 1–13. 10.1186/s12866-016-0653-3. [PubMed: 26728027]

Rupnik M, Wilcox MH and Gerding DN (2009). “*Clostridium difficile* infection: new developments in epidemiology and pathogenesis.” Nature Reviews Microbiology. 7(7): 526–536. [PubMed: 19528959]

Schaeffler H and Breitrueck A (2018). “*Clostridium difficile*–from colonization to infection.” Frontiers in microbiology. 9: 646. 10.1038/nrmicro2164. [PubMed: 29692762]

Schechner V, Fallach N, Braun T, Temkin E and Carmeli Y (2021). “Antibiotic exposure and the risk of hospital-acquired diarrhoea and *Clostridioides difficile* infection: A cohort study.” Journal of Antimicrobial Chemotherapy. 76(8): 2182–2185. 10.1093/jac/dkab151. [PubMed: 33969419]

Seemann T (2014). “Prokka: rapid prokaryotic genome annotation.” Bioinformatics. 30(14): 2068–2069. 10.1093/bioinformatics/btu153. [PubMed: 24642063]

Seemann T (2016) Shovill Github https://github.com/tseemann/shovill.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Shehabi AA, Badran EF and Abu-Khader EN (2015). "*Clostridium difficile*: Infection, diagnosis and treatment with antimicrobial drugs: A review article." The International Arabic Journal of Antimicrobial Agents. 5(4). 10.3390/pathogens13020118.

Spigaglia P, Mastrantonio P and Barbanti F (2018). "Antibiotic resistances of *Clostridium difficile*." Updates on *Clostridium difficile* in Europe: Advances in Microbiology, Infectious Diseases and Public Health. 8: 137–159. 10.1007/978-3-319-72799-8_15.

Stone NE, Sidak-Loftis LC, Sahl JW, Vazquez AJ, Wiggins KB, Gillece JD, Hicks ND, Schupp JM, Busch JD and Keim P (2016). "More than 50% of *Clostridium difficile* isolates from pet dogs in Flagstaff, USA, carry toxigenic genotypes." PloS one. 11(10): e0164504. 10.1371/journal.pone.0164504. [PubMed: 27723795]

Storey JD (2003). "The positive false discovery rate: a Bayesian interpretation and the q-value." The annals of statistics. 31(6): 2013–2035. 10.1214/aos/1074290335.

Surawicz CM, Brandt LJ, Binion DG, Ananthakrishnan AN, Curry SR, Gilligan PH, McFarland LV, Mellow M and Zuckerbraun BS (2013). "Guidelines for diagnosis, treatment, and prevention of *Clostridium difficile* Infections." Official journal of the American College of Gastroenterology| ACG. 108(4): 478–498. 10.1038/ajg.2013.4.

Tonkin-Hill G, MacAlasdair N, Ruis C, Weimann A, Horesh G, Lees JA, Gladstone RA, Lo S, Beaudoin C and Floto RA (2020). "Producing polished prokaryotic pangenomes with the Panaroo pipeline." Genome biology. 21: 1–21. 10.1186/s13059-020-02090-4.

Vernikos G, Medini D, Riley DR and Tettelin H (2015). "Ten years of pan-genome analyses." Current opinion in microbiology. 23: 148–154. 10.1016/j.mib.2014.11.016. [PubMed: 25483351]

Von Wintersdorff CJ, Penders J, Van Niekerk JM, Mills ND, Majumder S, Van Alphen LB, Savelkoul PH and Wolffs PF (2016). "Dissemination of antimicrobial resistance in microbial ecosystems through horizontal gene transfer." Frontiers in microbiology. 7: 173. 10.3389/fmicb.2016.00173. [PubMed: 26925045]

Weber DJ, Anderson DJ, Sexton DJ and Rutala WA (2013). "Role of the environment in the transmission of *Clostridium difficile* in health care facilities." American journal of infection control. 41(5): S105–S110. 10.1016/j.ajic.2012.12.009. [PubMed: 23622740]

Yang T, Zhong J, Zhang J, Li C, Yu X, Xiao J, Jia X, Ding N, Ma G and Wang G (2018). "Pan-genomic study of *Mycobacterium tuberculosis* reflecting the primary/secondary genes, generality/individuality, and the interconversion through copy number variations." Frontiers in microbiology. 9: 1886. 10.3389/fmicb.2018.01886. [PubMed: 30177918]

Zakham F, Sironen T, Vapalahti O and Kant R (2021). "Pan and core genome analysis of 183 *Mycobacterium tuberculosis* strains revealed a high inter-species diversity among the human adapted strains." Antibiotics. 10(5): 500. 10.3390/antibiotics10050500. [PubMed: 33924811]

Zhang L, Dong D, Jiang C, Li Z, Wang X and Peng Y (2015). "Insight into alteration of gut microbiota in *Clostridium difficile* infection and asymptomatic *C. difficile* colonization." Anaerobe 34: 1–7. 10.1016/j.anaerobe.2015.03.008. [PubMed: 25817005]

Zhou Y, Jiang D, Yao X, Luo Y, Yang Z, Ren M, Zhang G, Yu Y, Lu A and Wang Y (2023). "Pan-genome wide association study of *Glaesserella parasuis* highlights genes associated with virulence and biofilm formation." Frontiers in Microbiology. 14: 1160433. 10.3389/fmicb.2023.1160433. [PubMed: 37138622]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

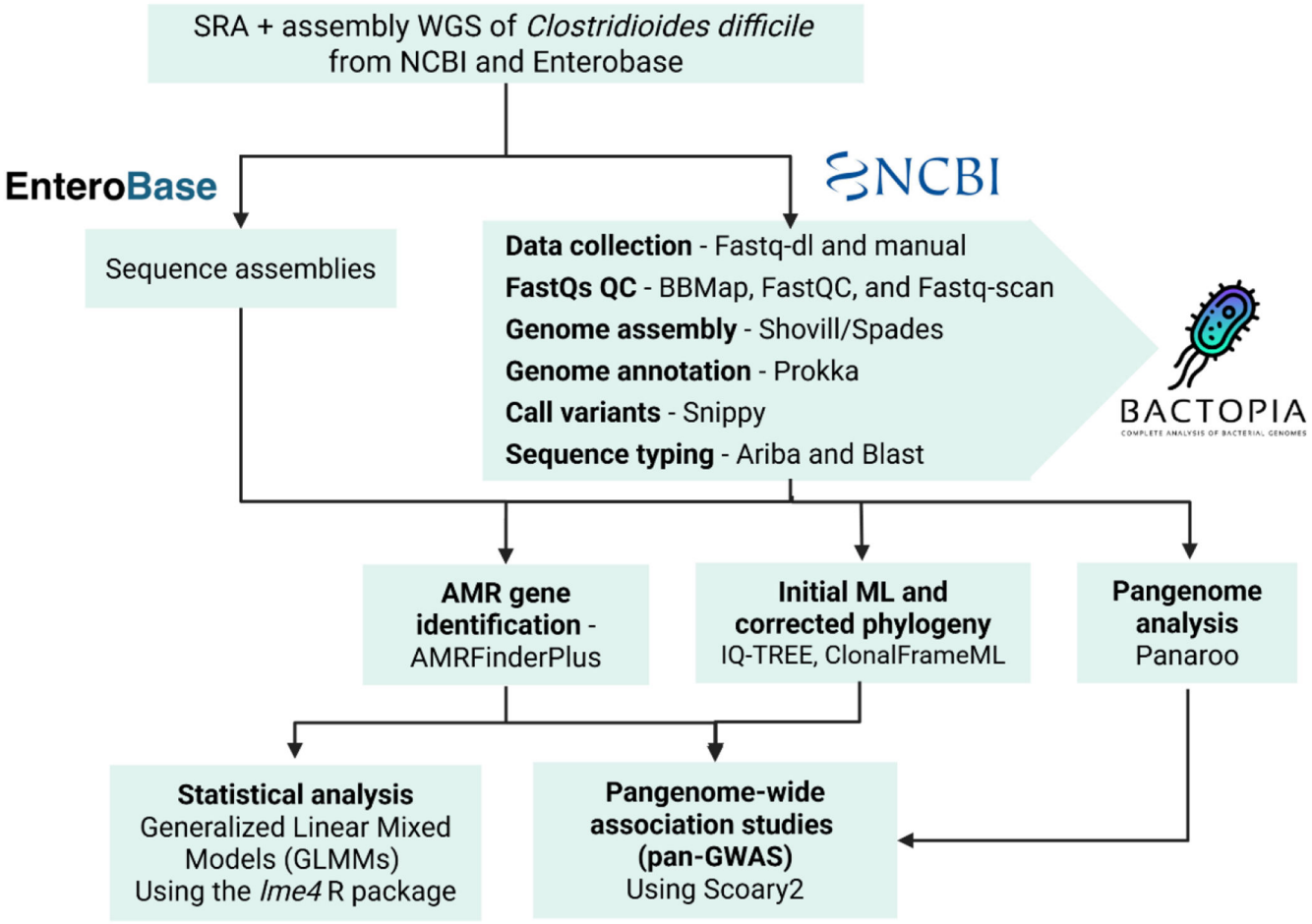


Figure 1.
Complete workflow for performing pangenome-GWAS analysis and visualization of *C. difficile* isolates.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

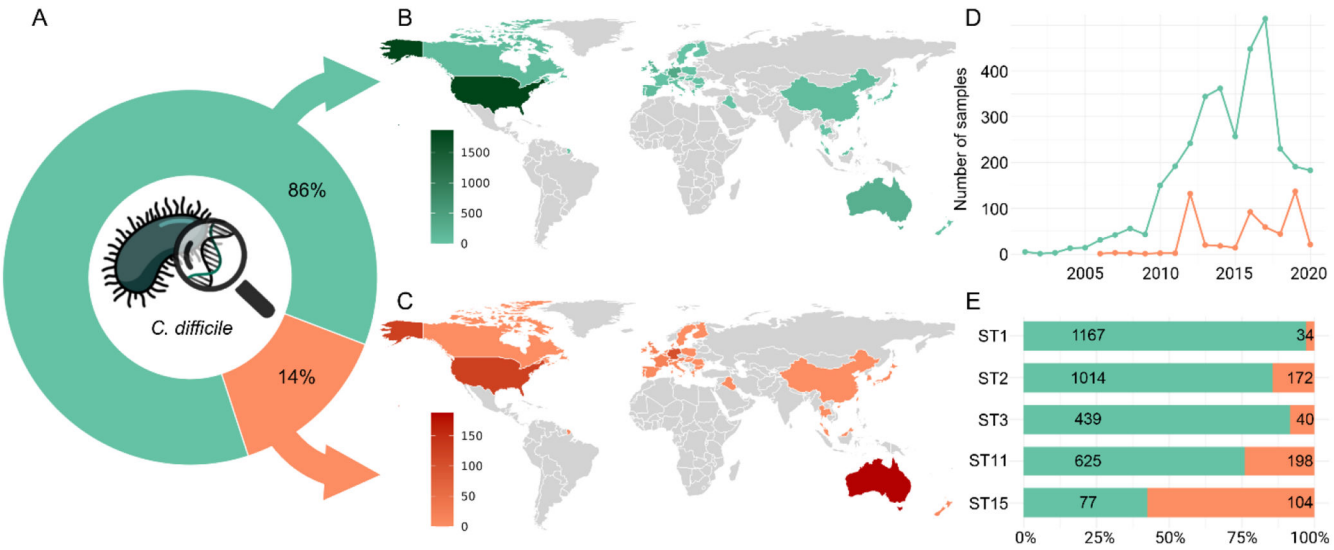


Figure 2.

Data description of the selected *C. difficile* sequences. The variation in green color represents the proportion of isolates that come from healthcare settings, while the variation in orange is based on the isolates that come from non-healthcare samples. A) Pie chart representing the overall distribution of 3,870 *C. difficile* isolates by setting. B) Global distribution of healthcare-associated samples, with countries shaded according to the number of samples collected. C) Global distribution of non-healthcare samples. D) Temporal distribution of samples from healthcare and external environments. E) Distribution of sequence types (STs) across healthcare and external environments (including number of samples inside each bar).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

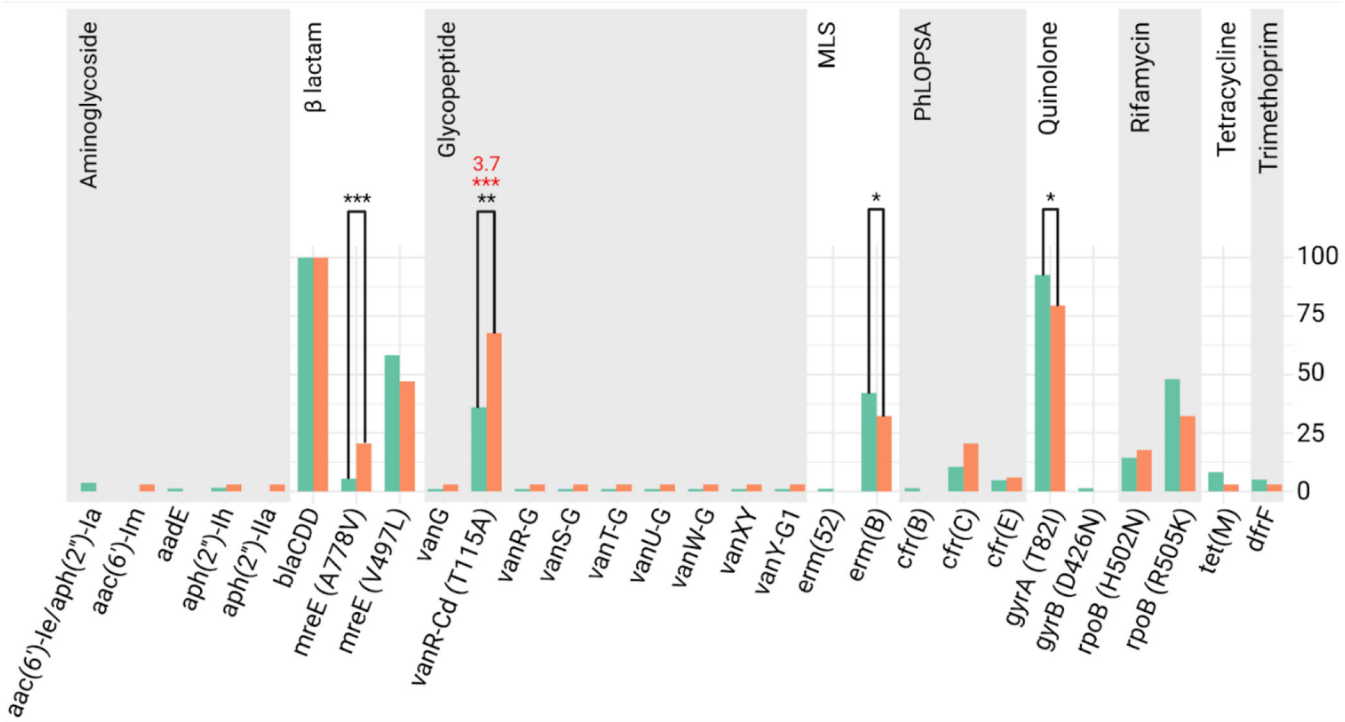


Figure 3.

Comparison of AMR determinant prevalence between *C. difficile* ST1 isolates from healthcare (green bars) and non-healthcare (orange bars) settings, grouped by AMR class. Each bar represents the proportion of isolates carrying the indicated determinant. Statistically significant differences between the two settings are marked by black asterisks when the results are based on generalized linear mixed models (GLMM) and by red asterisks when the results are referring to pangenome-wide association studies (pan-GWAS). Odds ratios for pan-GWAS associations are displayed in red numbers above the corresponding determinants, representing the strength of association for determinants more prevalent in one environment over the other. This figure displays the determinants with frequencies higher than 1%.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

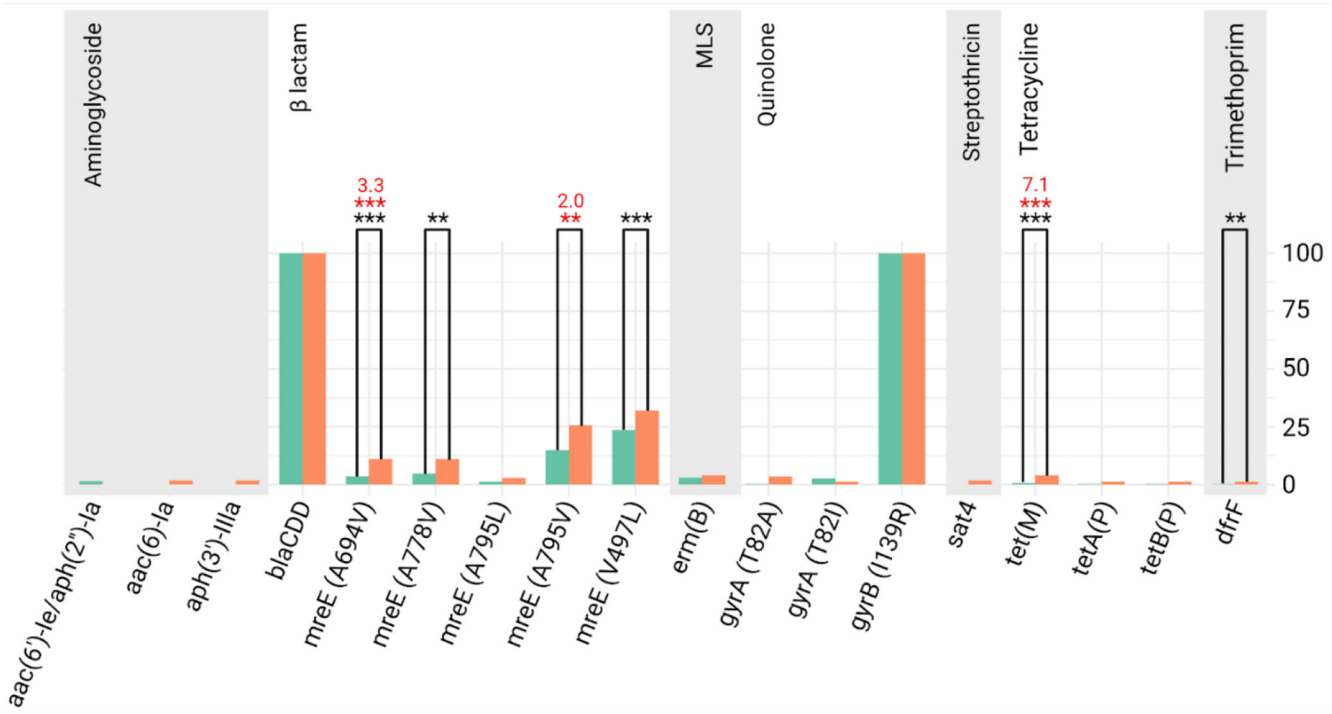


Figure 4.

Comparison of AMR determinant prevalence between *C. difficile* ST2 isolates from healthcare (green bars) and non-healthcare (orange bars) settings, grouped by AMR class. Each bar represents the proportion of isolates carrying the indicated determinant. Statistically significant differences between the two settings are marked by black asterisks when the results are based on generalized linear mixed models (GLMM) and by red asterisks when the results are referring to pangenome-wide association studies (pan-GWAS). Odds ratios for pan-GWAS associations are displayed in red numbers above the corresponding determinants, representing the strength of association for determinants more prevalent in one environment over the other. This figure displays the determinants with frequencies higher than 1%.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

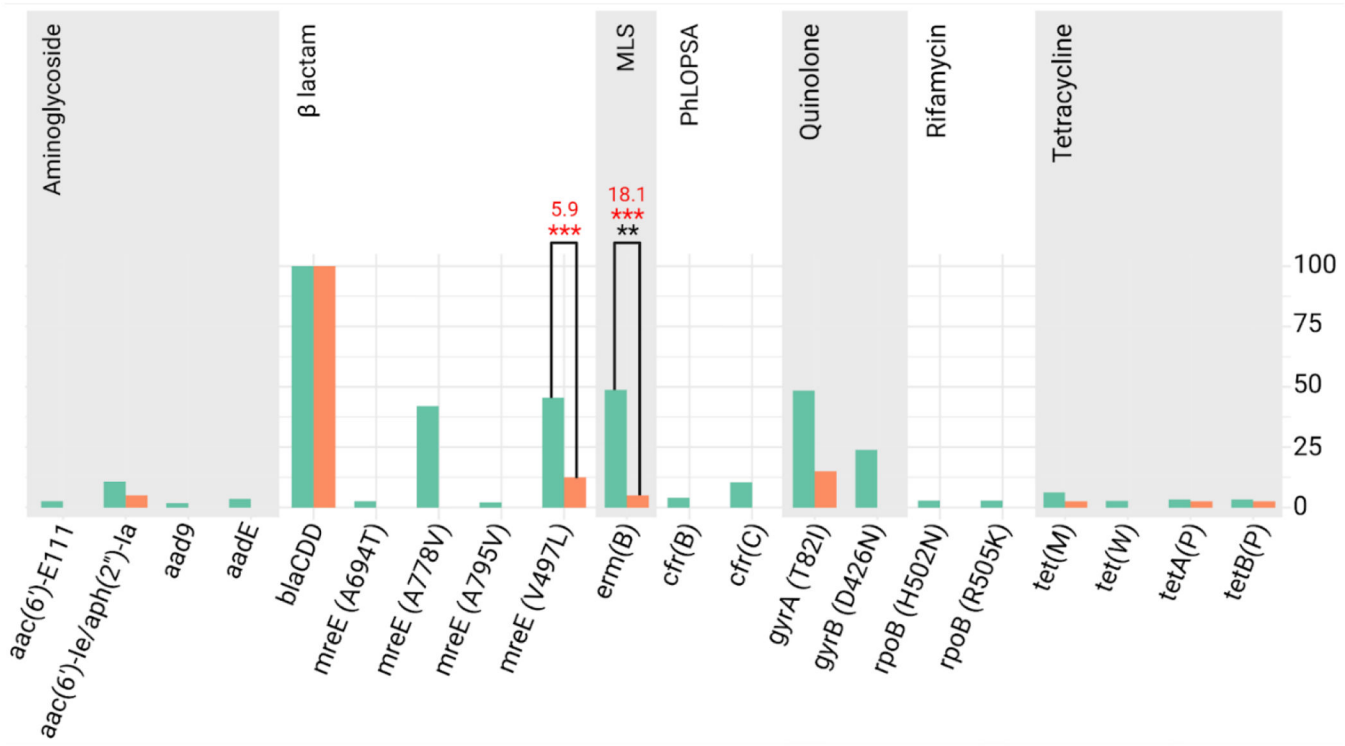


Figure 5.

Comparison of AMR determinant prevalence between *C. difficile* ST3 isolates from healthcare (green bars) and non-healthcare (orange bars) settings, grouped by AMR class. Each bar represents the proportion of isolates carrying the indicated determinant. Statistically significant differences between the two settings are marked by black asterisks when the results are based on generalized linear mixed models (GLMM) and by red asterisks when the results are referring to pangenome-wide association studies (pan-GWAS). Odds ratios for pan-GWAS associations are displayed in red numbers above the corresponding determinants, representing the strength of association for determinants more prevalent in one environment over the other. This figure displays the determinants with frequencies higher than 1%.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

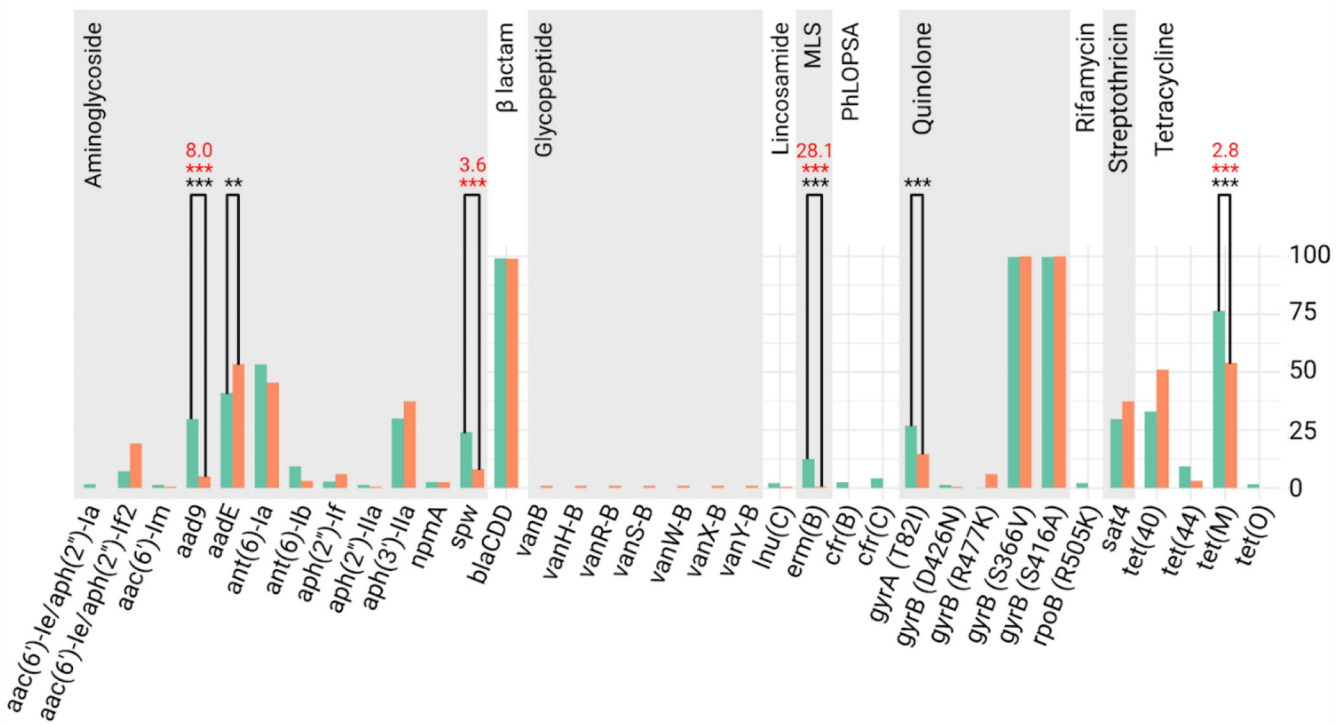


Figure 6.

Comparison of AMR determinant prevalence between *C. difficile* ST11 isolates from healthcare (green bars) and non-healthcare (orange bars) settings, grouped by AMR class. Each bar represents the proportion of isolates carrying the indicated determinant. Statistically significant differences between the two settings are marked by black asterisks when the results are based on generalized linear mixed models (GLMM) and by red asterisks when the results are referring to pangenome-wide association studies (pan-GWAS). Odds ratios for pan-GWAS associations are displayed in red numbers above the corresponding determinants, representing the strength of association for determinants more prevalent in one environment over the other. This figure displays the determinants with frequencies higher than 1%.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

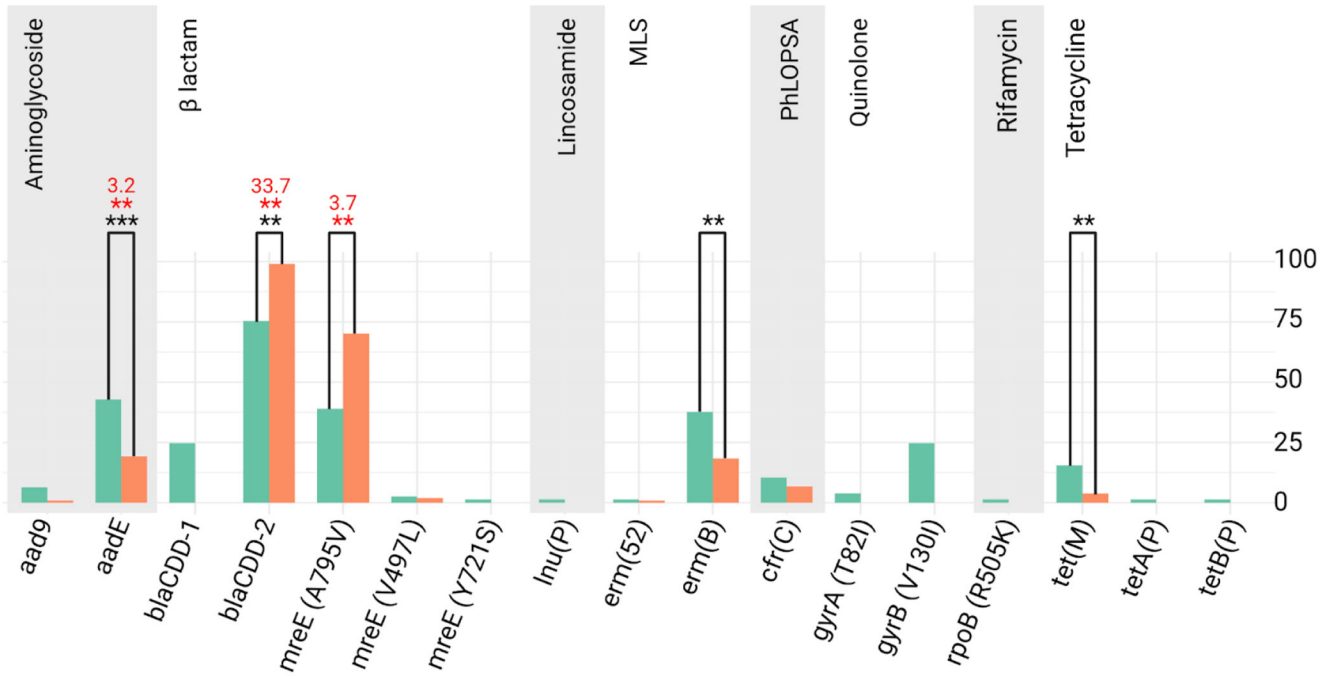


Figure 7.

Comparison of AMR determinant prevalence between *C. difficile* ST15 isolates from healthcare (green bars) and non-healthcare (orange bars) settings, grouped by AMR class. Each bar represents the proportion of isolates carrying the indicated determinant. Statistically significant differences between the two settings are marked by black asterisks when the results are based on generalized linear mixed models (GLMM) and by red asterisks when the results are referring to pangenome-wide association studies (pan-GWAS). Odds ratios for pan-GWAS associations are displayed in red numbers above the corresponding determinants, representing the strength of association for determinants more prevalent in one environment over the other. This figure displays the determinants with frequencies higher than 1%.