Progress in Microbes and Molecular Biology

HH PUBLISHER

Original Research Article

Phylogenomics, antimicrobial resistance, and mobile genetic elements in clinical *Clostridioides difficile* isolates from Hong Kong, China

Loh Teng-Hern Tan[1,2], Jodi Woan-Fei Law[1,2], Kah-Ooi Chua[3], Kok-Gan Chan[1,2,4], Rita WY Ng[5], Wing Shan Lee[5], Margaret Ip[5], Sunny Hei Wong[6,7,8*], Learn-Han Lee[1,2*]

Article History

Received: 09 July 2026

Received in Revised Form: 01 August 2026;

Accepted: 05 August 2026;

Available Online: 07 August 2026

[1]Centre for Molecular and Microbiome Sciences, Research Innovation Institute, School of Health and Life Sciences, University of Nottingham Ningbo China, Ningbo 315000, China; loh-teng-hern.tan@nottingham.edu.cn (LT-HT); jodi-woan-fei.law@nottingham.edu.cn (JW-FL)

[2]Microbiome and Bioresource Research Strength (MBRS), Jeffrey Cheah School of Medicine and Health Sciences, Monash University Malaysia, Bandar Sunway 47500, Selangor Darul Ehsan, Malaysia

[3]Centre for Research in Biotechnology for Agriculture (CEBAR), University of Malaya, Kuala Lumpur 50603, Malaysia; kahooi@um.edu.my (K-OC)

[4]Division of Genetics and Molecular Biology, Institute of Biological Sciences, Faculty of Science, University of Malaya, Kuala Lumpur 50603, Malaysia; kokgan@um.edu.my (K-GC)

[5]Department of Microbiology, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China; ritang@cuhk.edu.hk (RWYN); coral-lee@cuhk.edu.hk (WSL); margaretip@cuhk.edu.hk (MI)

[6]State Key Laboratory of Digestive Diseases, Department of Medicine and Therapeutics, Faculty of Medicine, The Chinese University of Hong Kong, Hong Kong SAR, China

[7]Centre for Microbiome Medicine, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore

[8]Department of Gastroenterology and Hepatology, Tan Tock Seng Hospital, Singapore

*Corresponding author: Learn-Han Lee; Centre for Molecular and Microbiome Sciences, Research Innovation Institute, School of Health and Life Sciences, University of Nottingham Ningbo China, Ningbo 315000, China; learn-han.lee@nottingham.edu.cn (L-HL); Sunny Hei Wong; Centre for Microbiome Medicine, Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore; sunny.wong@ntu.edu.sg (SHW)

Abstract: Whole-genome sequencing enables high-resolution characterisation of the evolutionary and accessory-genome diversity of *Clostridioides difficile* beyond conventional sequence typing. We analysed 118 clinical *C. difficile* genomes from Hong Kong using recombination-filtered phylogenomics, pangenome analysis, antimicrobial-resistance and virulence screening, mobile-element context analysis and functional annotation. The population comprised 24 sequence types and formed distinct phylogenetic clusters that were broadly consistent with multilocus sequence typing assignments. Pangenome analysis

identified 9,140 gene clusters and supported an open pangenome structure, indicating ongoing gene acquisition and loss within the population. Accessory-genome divergence increased with core-genome SNP distance (Spearman's $\rho = 0.437$, $p < 1 \times 10^{-5}$), although some closely related isolates retained substantial differences in gene content, suggesting recent accessory-genome turnover. Antimicrobial resistance determinants showed stronger associations with mobile genetic elements (MGE) than with virulence-associated loci. In particular, *erm(B)*, *AAC/APH*-family genes and *tet(M)* were frequently detected and commonly linked to insertion-sequence- or transposon-associated contexts, whereas major toxin loci showed limited evidence of local mobility. Functional annotation revealed broadly conserved functional profiles dominated by transcription-, metabolism-, and signal transduction-related categories, alongside significant sequence type-associated variation in KEGG pathway composition (PERMANOVA $R^2 = 0.434$, $p = 0.001$), with ST54 forming a distinct functional cluster. Collectively, these findings demonstrate that this Hong Kong *C. difficile* population is structured by stable evolutionary lineages while maintaining extensive accessory-genome diversity, and highlight the prominent role of MGE in shaping antimicrobial-resistance architecture relative to the largely conserved virulence repertoire.

Keywords: *Clostridioides difficile*; whole-genome sequencing; virulence; antimicrobial resistance; pangenome; mobile genetic elements; functional annotation

1. Introduction

Clostridioides difficile is a leading cause of healthcare-associated and antimicrobial-associated diarrhoeal disease worldwide and remains a major source of morbidity, mortality, and healthcare expenditure, particularly among hospitalised and elderly patients [1–3]. Disease severity ranges from mild self-limiting diarrhoea to fulminant colitis and recurrent infection, with outcomes influenced by host susceptibility, antimicrobial exposure, bacterial virulence, and transmission dynamics [1, 4]. This clinical burden sits within a wider antimicrobial-resistance landscape in which antimicrobial misuse, persistent drug-resistant pathogens and narrowing treatment options continue to increase the value of genome-informed bacterial surveillance [5]. The emergence and international dissemination of successful epidemic lineages have further highlighted the importance of understanding the evolutionary processes underlying *C. difficile* population structure and adaptation [4, 6].

The pathogenicity of *C. difficile* is primarily associated with the production of toxins A (*tcdA*) and B (*tcdB*), encoded within the chromosomal pathogenicity locus (PaLoc), while a subset of strains additionally carries the binary toxin locus (*CdtLoc*) [7]. Comparative genomic studies have demonstrated that both toxin-associated loci and antimicrobial resistance (AMR) determinants have complex evolutionary histories involving recombination, horizontal gene transfer, gene loss, and lineage-specific diversification [7, 8]. Consequently, the clinical and epidemiological characteristics of *C. difficile* populations cannot be fully understood through toxin profiling or AMR screening alone.

Multilocus sequence typing (MLST) has been widely used to classify *C. difficile* isolates into sequence types (STs) and major evolutionary clades, providing a standardised framework for global surveillance and inter-study comparison [9, 10]. However, MLST interrogates only a small fraction of the genome and therefore has limited resolution for investigating recent transmission, recombination, accessory-genome variation, and mobile genetic elements (MGE) [11, 12]. Whole-genome sequencing (WGS) has transformed *C. difficile* epidemiology by enabling high-resolution reconstruction of phylogenetic relationships, identification of transmission clusters, characterisation of AMR determinants, and investigation of horizontal gene transfer events [4, 11, 12].

Genome-scale analyses have revealed that *C. difficile* possesses a highly dynamic and mosaic genome shaped by recombination and extensive accessory-gene exchange [8]. MGE, including insertion sequences, conjugative transposons, integrative conjugative elements, plasmids, and bacteriophages, contribute substantially to genomic diversification and facilitate the dissemination of AMR determinants such as *erm(B)* and *tet(M)* [13, 14]. At the same time, comparative genomic studies have demonstrated that closely related isolates may differ considerably in accessory-gene content despite sharing highly similar core genomes, emphasising the importance of integrating pangenome and phylogenomic analyses when interpreting population structure [8].

Recent advances in bacterial genomics have expanded the scope of WGS-based investigations beyond phylogenetic reconstruction [15]. Pangenome analysis can distinguish conserved and variable components of the genome, while functional annotation enables interpretation of gene-content variation in terms of biological pathways and ecological adaptation [16–18]. Similarly, modern tools for mobile genetic element prediction allow systematic investigation of plasmid-, prophage-, and insertion-sequence associated genomic regions, providing insight into potential mechanisms of resistance dissemination and genome evolution [19, 20]. Collectively, these approaches facilitate a more comprehensive understanding of how core-genome evolution, accessory-genome dynamics, and mobile-element activity interact within bacterial populations.

Despite the increasing application of WGS to *C. difficile* surveillance globally, comprehensive genomic characterisation of isolates from Asia remains limited [21–23]. In particular, few studies have integrated lineage assignment, recombination-aware phylogenomics, pangenome analysis, AMR profiling, mobile genetic element prediction, and functional annotation. In other opportunistic bacterial pathogens, integrated genomic and resistome analyses have similarly shown how taxonomic confirmation, phylogenetic placement and resistance-gene profiling can clarify the clinical relevance of draft-genome data [24]. Such integrated analyses are important because they provide complementary perspectives on population structure and can generate testable hypotheses regarding the evolution, dissemination, and ecological adaptation of clinically relevant lineages [25]. Public bioinformatic platforms such as Galaxy can support reproducible analysis by preserving tool versions, parameters and workflow provenance, which is particularly useful for multi-step microbial genomics studies [26].

In this study, we performed an integrated comparative genomic analysis of 118 *C. difficile* genome assemblies from Hong Kong, China. We aimed to determine whether this regional population shows lineage-structured genomic diversity, whether AMR determinants and virulence-associated loci differ in their local MGE-neighbourhood evidence and whether functional annotation reveals sequence type (ST)-associated differences beyond MLST assignment. To address these aims, we combined MLST-based lineage assignment, recombination-filtered pangenomics, pangenome analysis, AMR and virulence gene screening, MGE-context analysis, insertion-sequence (IS) prediction, plasmid and viral region prediction and eggNOG/KEGG functional annotation. This integrated design was intended to provide a molecular genomic baseline for Hong Kong *C. difficile* isolates and to identify candidate resistance and functional patterns for future clinical and experimental validation.

2. Materials and Methods

2.1. Whole-Genome Sequencing and Assembly Annotation

A total of 118 *Clostridioides difficile* genome assemblies associated with NCBI BioProject PRJNA1063910 were included in this study. Whole-genome sequencing was performed using the DNBSEQ™ sequencing platform, and genomes were assembled de novo using SPAdes v3.13.0. Genome assemblies were annotated using Prokka v1.14.6 to generate standardised gene predictions and GFF3 annotation files for downstream analyses [27].

2.2. Multilocus Sequence Typing and Population Structure Analysis

STs were assigned using the seven-locus *C. difficile* MLST scheme implemented through the PubMLST framework [9, 10]. The Galaxy MLST tool was run using minimum DNA identity and coverage thresholds of 95% and 10%, respectively.

For phylogenomic reconstruction, reference-based variant calling was performed against the *C. difficile* CD630 chromosome (NC_009089.1) and plasmid (NC_008226.2) using Snippy [28]. Core-genome SNP alignments were extracted with SNP-sites [29], and recombinant regions were identified and removed using Gubbins [30].

Maximum-likelihood phylogenies were reconstructed using IQ-TREE v2.4.0 [31]. The optimal nucleotide substitution model was selected using ModelFinder [32], and branch support was assessed with 1,000 ultrafast bootstrap and 1,000 SH-aLRT replicates. The final phylogeny was pruned to the 118 isolates and annotated with ST, AMR determinants and virulence factors in iTOL (https://itol.embl.de/) [33]. The Snippy-core and Gubbins alignments contained the 118 isolates plus the CD630 reference sequence. For visualisation, the final tree was pruned to the 118 clinical isolates.

Galaxy wrapper versions, database releases, and analysis parameters are summarised in the supplementary material (Table S1).

2.3. Pangenome Analysis

Prokka-annotated GFF3 files were analysed using Roary [18] with a minimum BLASTP identity threshold of 95%, a core-gene threshold of 99%, an MCL inflation value of 1.5, and paralog splitting enabled. Gene clusters were classified as core, soft-core, shell, or cloud according to Roary frequency categories. Pangenome accumulation curves were generated from the Roary gene presence–absence matrix using 200 random genome-order permutations. Pangenome openness was estimated by fitting a Heaps-law power function:

$$n = \kappa N^{\gamma}$$

where n is the pangenome size, N is the number of genomes used, κ and γ are the fitting parameters.

To evaluate relationships between gene-content variation and core-genome evolution, pairwise accessory-genome distances were calculated from gene presence-absence profiles and compared with recombination-filtered SNP distances using Spearman rank correlation.

2.4. Antimicrobial Resistance and Virulence Gene Identification

AMR determinants were identified using AMRFinderPlus [34] and ABRicate [35]. AMRFinderPlus analyses employed database release amrfinderplus_V3.12_2024-05-02.2 with nucleotide sequence input and the *C. difficile* organism setting enabled.

ABRicate searches were performed independently against the CARD [36], ARG-ANNOT [37], NCBI AMR, and ResFinder [38] databases using minimum identity and coverage thresholds of 90%. Virulence-associated genes were screened using the Virulence Factor Database (VFDB) [39].

To minimise redundancy among database outputs, coordinate-bearing ABRicate hits were harmonised and deduplicated based on isolate, contig, gene annotation, and genomic coordinate overlap prior to downstream analyses.

2.5. Mobile Genetic Element Context Analysis

Potential mobility of AMR and virulence loci was evaluated by examining local genomic context. Coordinate-bearing ABRicate hits were first harmonised by gene name and deduplicated by isolate, contig, canonical gene label, and coordinate overlap. For each deduplicated locus, Prokka annotations within 10 kb upstream or downstream on the same contig were screened for terms indicating transposases, insertion sequences, integrases, excisionases, relaxases, conjugation or mobilisation proteins, plasmid-associated annotations, and phage or prophage-associated annotations. [40–42]. Distances were calculated edge-to-edge between the ABRicate locus and the Prokka feature, with overlapping intervals assigned a distance of 0 bp.

Evidence was classified as high-confidence local MGE-neighbourhood evidence when a strong MGE-associated annotation was located within 0–2 kb of the locus, moderate-confidence when a strong marker occurred within >2–10 kb, weak or indirect evidence when only a generic recombinase/resolvase annotation occurred within 10 kb. Additional MGE evidence came from geNomad [19] viral or plasmid-region predictions, ISEScan [20] insertion sequence predictions, and PlasmidFinder [43] replicon calls. Positive hits were interpreted as evidence of known plasmid replication systems but not as confirmation of complete plasmid reconstruction.

2.6. Functional Annotation and Lineage Association Analysis

Predicted protein sequences were functionally annotated using eggNOG-mapper [16] against the eggNOG v5.0.2 database [17] with DIAMOND-based searches [44]. KEGG Orthology (KO), KEGG pathway, KEGG module, and COG functional category assignments were extracted from annotation outputs [45]. Annotation coverage was summarised for each isolate as the proportion of predicted proteins assigned to KO, pathway and module categories. COG functional composition was summarised as the total number of annotated gene hits per category across the dataset.

KEGG pathway burden was calculated as the number of genes assigned to each pathway per isolate and converted to within-isolate relative pathway profiles to account for differences in annotation depth. For visualisation of lineage-associated functional variation, mean pathway profiles were calculated for major ST groups and standardised using row-wise z-scores. Selected pathways with biological relevance to metabolism, transport, signalling, motility and AMR were displayed as heatmaps.

Functional similarities among isolates were assessed using Bray-Curtis dissimilarities calculated from relative KEGG pathway profiles, followed by principal coordinates analysis (PCoA). Differences in functional composition among ST groups were evaluated using permutational multivariate analysis of variance (PERMANOVA) with 999 permutations.

3. Results

3.1. Recombination-filtered phylogeny resolved lineage structure and marker distribution

MLST analysis identified 24 STs, indicating substantial lineage diversity within the collection. Six dominant STs accounted for 72.9% (86/118) of isolates, comprising ST3 (20/118, 16.9%), ST8 (20/118, 16.9%), ST2 (16/118, 13.6%), ST54 (11/118, 9.3%), ST42 (10/118, 8.5%), and ST102 (9/118, 7.6%). Eight STs were represented by a single isolate, highlighting the presence of less common lineages within the population. Isolates from the same ST generally formed coherent phylogenetic clusters, indicating strong concordance between seven-locus typing and genome-wide population structure.

Despite the diversity of STs, the collection was strongly dominated by MLST clade 1, which comprised 112 of 118 isolates (94.9%). Clade 4 accounted for four isolates (3.4%),

whereas only two isolates (1.7%) belonged to clade 5. No representatives of clades 2 or 3 were detected. Thus, most genomic diversity within the collection reflected diversification within clade 1 rather than contributions from multiple deeply separated evolutionary clades.

The recombination-filtered phylogeny was consistent with MLST-based lineage assignments and demonstrated clear separation of the major clades (Figure 1). The two clade 5 isolates (HK69 and HK78; ST11) formed a distinct branch characterised by long evolutionary distances from the dominant clade 1 population, whereas clade 4 isolates occupied an intermediate position. These observations indicate that although this Hong Kong collection was overwhelmingly composed of clade 1 isolates, it nevertheless contained representatives of evolutionarily divergent lineages.

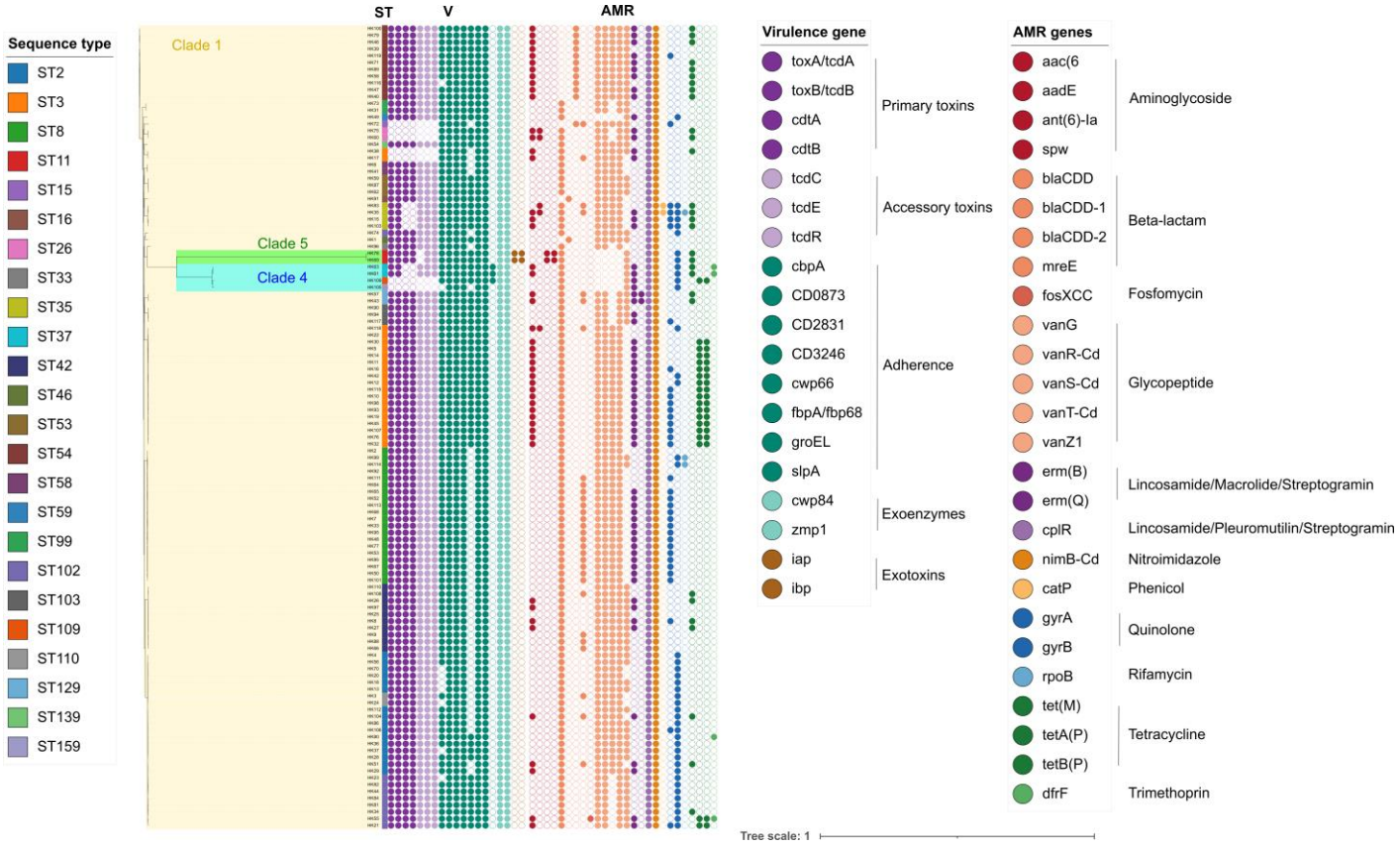

Figure 1. Recombination-filtered phylogenomic structure of 118 *C. difficile* isolates. Maximum-likelihood phylogeny reconstructed from recombination-filtered core-genome SNPs after Gubbins filtering and IQ-TREE inference. Concentric annotation tracks indicate ST, virulence-associated genes and AMR determinants. Major ST lineages form distinct phylogenetic clusters, demonstrating substantial population structure across the study collection.

The recombination-filtered analysis provided quantitative support for the broad lineage structure. The Gubbins-filtered alignment contained 162,002 polymorphic positions before pruning of the reference sequence for visualisation. Across isolate pairs, the median recombination-filtered distance was 7,179 SNPs (range, 0-98,744). Same-ST pairs had a substantially lower median filtered distance than different-ST pairs (50 versus 7,297 SNPs), supporting concordance between MLST and genome-wide structure. Recombination filtering removed a median of 4,929 pairwise SNP differences, corresponding to approximately 40.0%

of raw pairwise differences. Gubbins identified 2,574 branch-associated recombination events, with a median event length of 893 bp. These results indicate that recombination was an important feature of the dataset and support the use of recombination-filtered distances for phylogenomic interpretation.

3.2. Antimicrobial resistance and virulence-associated gene profiles

AMRFinderPlus identified 1,732 detection records corresponding to 33 unique feature names. Individual isolates carried a median of 15 resistance-associated features (range, 6–21). Several genes were ubiquitous or near-ubiquitous across the collection, including *cplR* and *nimB-Cd* (118/118), *vanG, vanR-Cd*, and *vanT-Cd* (111/118 each), *blaCDD* (104/118), and *vanS-Cd* (93/118). Among the more variably distributed resistance determinants, *erm(B)* was the most prevalent gene, occurring in 64 isolates (54.2%), followed by *aac(6')-Ie/aph(2'')-Ia* in 48 isolates (40.7%), *tet(M)* in 29 isolates (24.6%), and paired *tetA(P)*/*tetB(P)* loci in 19 isolates (16.1%). Less common determinants included *aadE* (5 isolates), *dfrF* (4 isolates), phenicol-resistance genes (2 isolates), and a single fosfomycin-associated gene. These genotypic findings indicate resistance potential, but they do not establish phenotypic resistance without antimicrobial susceptibility testing.

Virulence-associated genes were detected in the majority of isolates. AMRFinderPlus detected *tcdB*, *tcdE*, and *tcdR* in 111 isolates (94.1%), *tcdC* in 109 isolates (92.4%), and *tcdA* in 108 isolates (91.5%). Additional virulence-associated genes, including *CD0873, CD2831, CD3246, fbpA/fbp68, groEL*, *cwp84* and *zmp1*, were also highly prevalent across the collection. Screening against VFDB identified the binary toxin genes *cdtA* and *cdtB* in 103 isolates (87.3%), indicating that binary toxin loci were present in a substantial subset of the population.

A notable discrepancy was observed between VFDB and AMRFinderPlus outputs for several toxin-associated loci. Initial VFDB screening with a 90% coverage threshold detected *toxA* in only 3/118 isolates, whereas AMRFinderPlus identified the corresponding toxin gene (*tcdA*) in 108 isolates. Given the unexpectedly large difference, VFDB screening was repeated using a less stringent coverage threshold. Inspection of the VFDB output showed that many *toxA* matches exhibited high sequence identity but failed the initial coverage criterion. Similar differences were observed for *cdtA* and *cdtB*, with some loci detected by VFDB but not AMRFinderPlus and vice versa. These observations highlight the influence of database composition, reference-sequence selection, gene nomenclature, and detection thresholds on virulence-gene prevalence estimates.

Overall, virulence-associated loci showed a more uniform distribution across the collection than resistance-associated determinants (Figure 1). Major toxin genes were detected in more than 90% of isolates, whereas determinants such as *erm(B)*, *aac(6')-Ie/aph(2'')-Ia* (*AAC/APH*), *tet(M)*, and *tetA(P)/tetB(P)* were present only in subsets of the

population. This contrast in distribution patterns provided a basis for subsequent analyses of accessory-genome variation and mobile-element context.

3.3. An open pangenome revealed extensive accessory diversity

Pangenome analysis of the 118 *Clostridioides difficile* genomes identified 9,140 orthologous gene clusters (Figure 2). The pangenome accumulation curve continued to increase with the sequential addition of genomes and did not approach saturation (Figure 2A). Fitting the data to Heaps' law yielded an exponent of $\gamma = 0.206$, consistent with an open pangenome and indicating that additional genomes would likely contribute novel gene content. The complete Roary pangenome output is provided in the supplementary file (Figure S1).

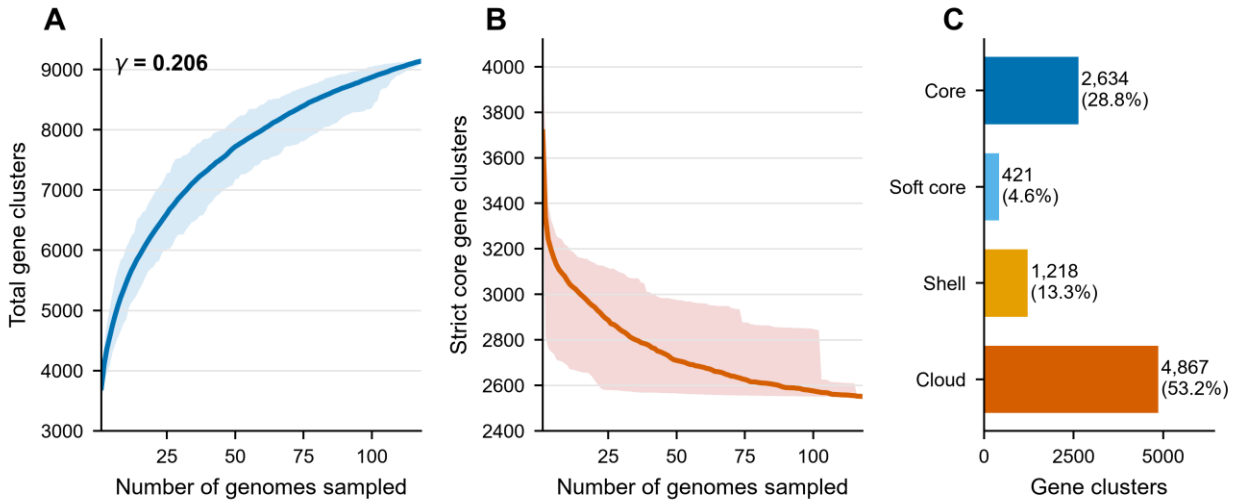

Figure 2. Pangenome architecture of 118 *C. difficile* genomes. (A) Pangenome accumulation curve showing the increase in total gene clusters as genomes were added. The fitted Heaps' law parameter gamma, $\gamma = 0.206$, indicates an open pangenome. Shaded regions represent variation across random genome permutations. (B) Strict core-genome decay curve showing the reduction in core-gene clusters as additional genomes were incorporated. (C) Distribution of gene clusters among pangenome classes.

The strict core-genome curve decreased with additional genomes and approached approximately 2,600 clusters (Figure 2B). Frequency-class analysis showed that 2,634 clusters were core genes (28.8%), 421 were soft-core genes (4.6%), 1,218 were shell genes (13.3%), and 4,867 were cloud genes (53.2%) (Figure 2C). Consequently, more than 70% of all gene clusters belonged to the accessory genome, being absent from at least one isolate. The predominance of cloud genes highlights substantial gene-content variability within the population and suggests that gene gain and loss contribute substantially to genomic diversification.

3.4. Accessory gene diversity was shaped by both lineage and gene gain/loss

Given the extensive accessory-genome diversity identified in the pangenome analysis, we next examined whether variation in gene content tracked evolutionary relationships inferred from the core genome. Pairwise accessory-gene distances were therefore compared with recombination-filtered pairwise SNP distances across 4,582 pairwise isolate comparisons.

Accessory-genome distance showed a moderate positive correlation with recombination-filtered SNP distance (Spearman's ρ = 0.437, $p = 1 \times 10^{-5}$), indicating that gene-content variation was partly structured by evolutionary relationships (Figure 3). Comparisons between isolates belonging to the same ST generally exhibited low SNP and accessory-gene distances, whereas comparisons between different STs occupied the upper range of both measures. Consistent with this pattern, median pairwise accessory-gene differences were substantially lower within STs (232 gene clusters) than between STs (751 gene clusters).

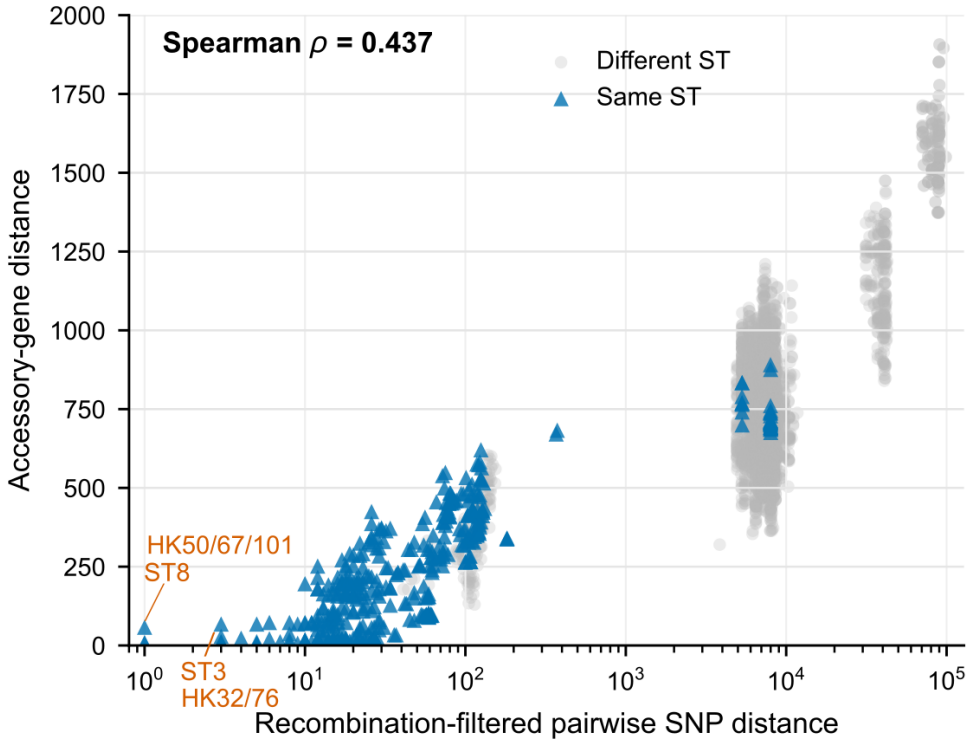

Figure 3. Relationship between core-genome divergence and accessory-genome variation. Pairwise accessory-gene distances were plotted against recombination-filtered pairwise SNP distances. Blue triangles represent same-ST comparisons, and grey circles represent different-ST comparisons.

Despite this overall relationship, considerable scatter was observed among isolate pairs with similar SNP distances, indicating that accessory-genome evolution is not determined solely by vertical inheritance. Instead, gain and loss of accessory genes contribute

additional diversity beyond that captured by core-genome phylogeny. Several closely related isolate groups illustrated this pattern. For example, the ST8 cluster comprising isolates HK50, HK67 and HK101 differed by only 0 to 1 recombination-filtered SNPs but varied by up to 69 accessory-gene clusters. A similar pattern was observed for the ST3 isolate pair HK32 and HK76. These observations suggest recent acquisition or loss of accessory elements, potentially involving plasmids, prophages, insertion sequences, or other MGE.

3.5. Resistance loci showed stronger mobile-element context than virulence loci

To evaluate whether resistance and virulence determinants were associated with MGE, we integrated local Prokka-context screening, ISEScan insertion-sequence prediction, geNomad viral/plasmid overlap analysis, and PlasmidFinder replicon calls. The mobility heatmap separated these evidence sources rather than collapsing them into a single mobility score, allowing direct comparison of the type and strength of MGE evidence for each locus (Figure 4).

MGE-context signals were concentrated around selected AMR genes. *erm(B)* showed both high- and moderate-confidence local evidence: 42/64 loci (65.6%) had a Prokka MGE marker within 2 kb, 19/64 loci (29.7%) had a marker at 2-10 kb, and ISEScan identified IS elements within 2 kb of 32/64 loci (50.0%). The bifunctional aminoglycoside determinant *AAC/APH* showed a similar but more strongly local pattern, with 37/48 loci (77.1%) having a Prokka MGE marker within 2 kb and 17/48 loci (35.4%) having an ISEScan IS within 2 kb. In contrast, *tet(M)* was mainly linked to more distant MGE context, with 17/29 loci (58.6%) associated with a Prokka MGE marker at 2-10 kb and 6/29 loci (20.7%) associated only with generic recombinase/resolvase evidence.

The MGE-type summary clarified the biological identity of these signals. *Tn916*-like annotations were the most common Prokka-context signal, occurring in 49 isolates and 57 loci, consistent with the known association of tetracycline resistance and conjugative transposon-like elements. IS*120*-family transposases were detected within 2 kb in 41 isolates and 45 loci, while IS*256*-family transposases were detected within 2 kb in 36 isolates and 36 loci. ISEScan independently supported a major insertion-sequence component, especially IS*3*-family signals, which were detected within 10 kb in 37 isolates and within 2 kb in 33 isolates.

PlasmidFinder replicons were common but were treated as indirect evidence because the available analysis did not establish the same-contig or same-element localisation between replicons and resistance genes. *rep1* was detected in 42 isolates and *repUS43* in 15 isolates. These calls are therefore best interpreted as replicon-associated genomic backgrounds rather than proof that specific AMR genes were plasmid-borne.

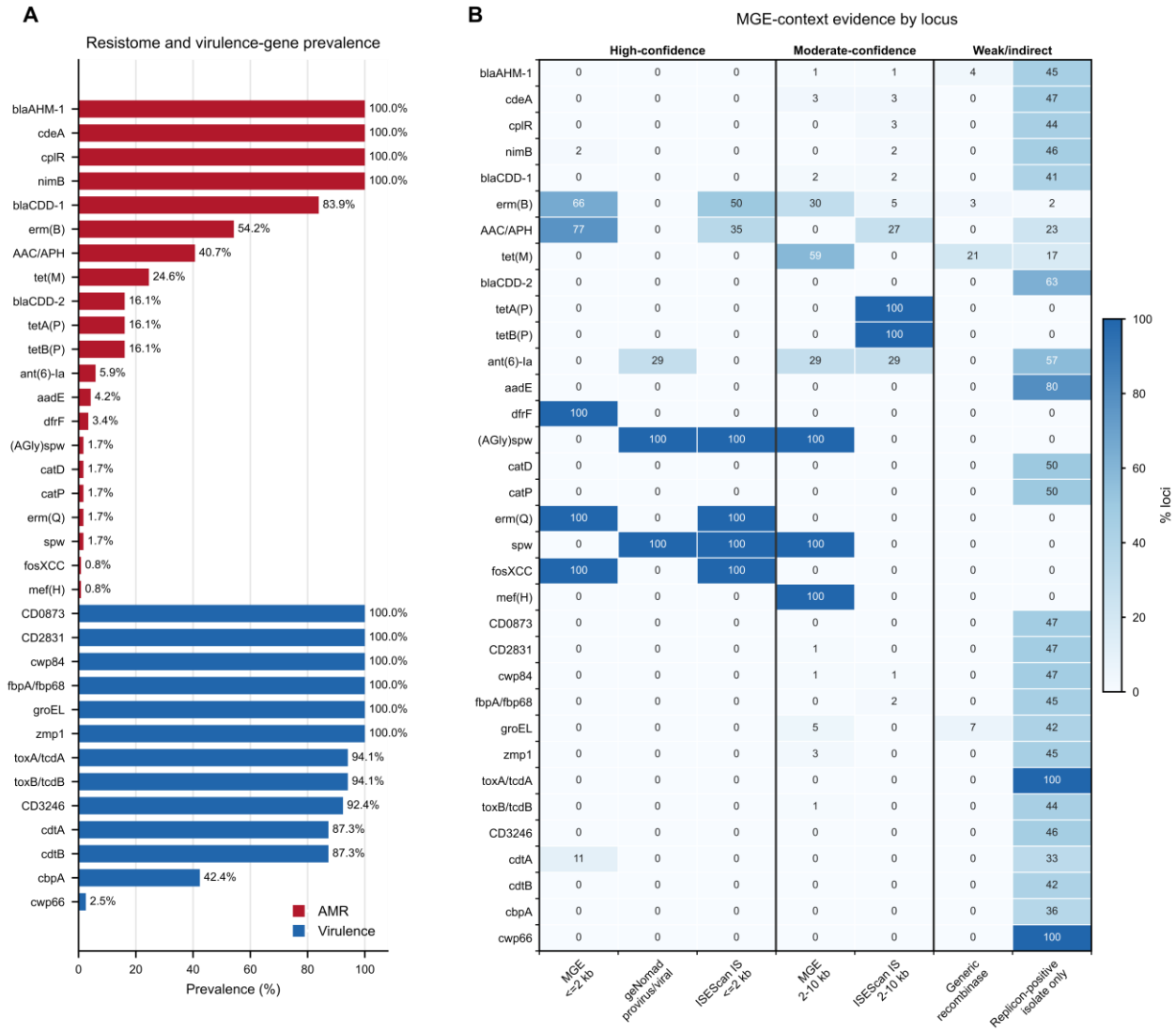

Figure 4. Mobile genetic element context of AMR and virulence loci. (A) Prevalence of AMR and virulence-associated genes among 118 *C. difficile* isolates. (B) The percentage of coordinate-deduplicated loci with individual MGE-context evidence types. Three-level mobile-element context evidence by locus. High-confidence evidence includes a Prokka-annotated MGE marker within 2 kb, geNomad provirus/viral coordinate overlap, or an ISEScan insertion sequence within 2 kb. Moderate-confidence evidence includes a Prokka MGE marker at >2-10 kb or an ISEScan IS at >2-10 kb. Weak/indirect evidence includes generic recombinase/resolvase annotations within 10 kb or PlasmidFinder replicon-positive isolate background only. Values represent the percentage of loci for each gene.

Finally, to integrate lineage, resistance-gene identity, and MGE context, we summarised AMR-associated flows using a Sankey diagram (Figure 5). This representation showed that the major AMR-MGE signal was not randomly distributed across the population. ST3 contributed the largest number of AMR loci in the Sankey analysis, with 120 loci across 20 isolates, followed by ST54 with 56 loci across 11 isolates and other STs collectively contributing 69 loci. The dominant AMR nodes were *AAC/APH* and *erm(B)*, contributing 113 and 112 AMR loci, respectively. These flowed mainly into IS*120*/IS*3* transposase, IS*256* transposase, and *Tn916*-like transposon categories. Specifically, *AAC/APH* linked strongly

to IS*120*/IS*3* transposase and IS*256* transposase, while *erm(B)* linked to IS*120*/IS*3*, IS*256*, and *Tn916*-like contexts. *tet(M)* flowed primarily into the *Tn916*-like transposon category. Together, the heatmap, MGE-type summary, and Sankey diagram support a mobilome architecture in which major AMR determinants are structured by lineage and by recurrent insertion-sequence/transposon-associated contexts, whereas virulence loci show comparatively limited direct local MGE association.

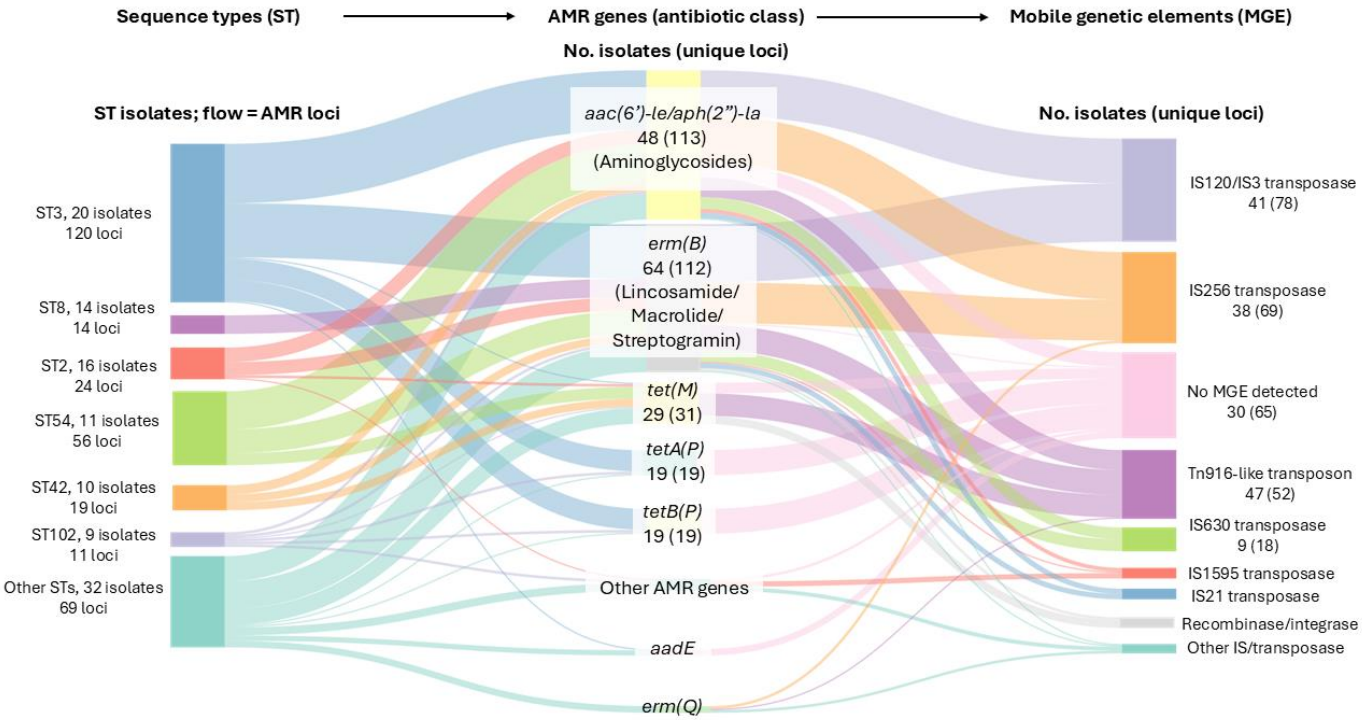

Figure 5. Lineage-resolved relationships among STs, AMR genes and MGE**.** Sankey diagram showing the relationship between ST, AMR gene identity, and nearby MGE category. Left nodes show major ST groups, middle nodes show AMR genes, and right nodes show MGE categories detected in the mobile-neighbourhood analysis. Flow thickness represents the number of AMR loci. ST labels report the number of isolates in the full collection and the number of AMR loci contributing to the Sankey flow. AMR gene and MGE labels show the number of isolates followed by the number of unique AMR loci in parentheses. The six most frequent STs were displayed individually; the remaining STs were grouped as "Other STs". Lower-frequency AMR genes were grouped as "Other AMR genes". MGE categories were derived from nearby mobile-feature annotations within the mobile-context analysis. The Sankey diagram was created using SankeyMATIC (http://sankeymatic.com).

3.6. Functional annotation showed stable coverage and ST-associated profile shifts

Functional annotation was summarised from eggNOG-mapper outputs for 414,494 predicted proteins across the 118 genomes (Figure 6). Annotation coverage was consistent across isolates, with mean assignment rates of 67.95% for KEGG Orthology terms, 41.49% for KEGG pathways and 29.72% for KEGG modules (Figure 6A). Per-isolate medians were 1,966.5 unique KOs, 208 KEGG pathways and 294 KEGG modules.

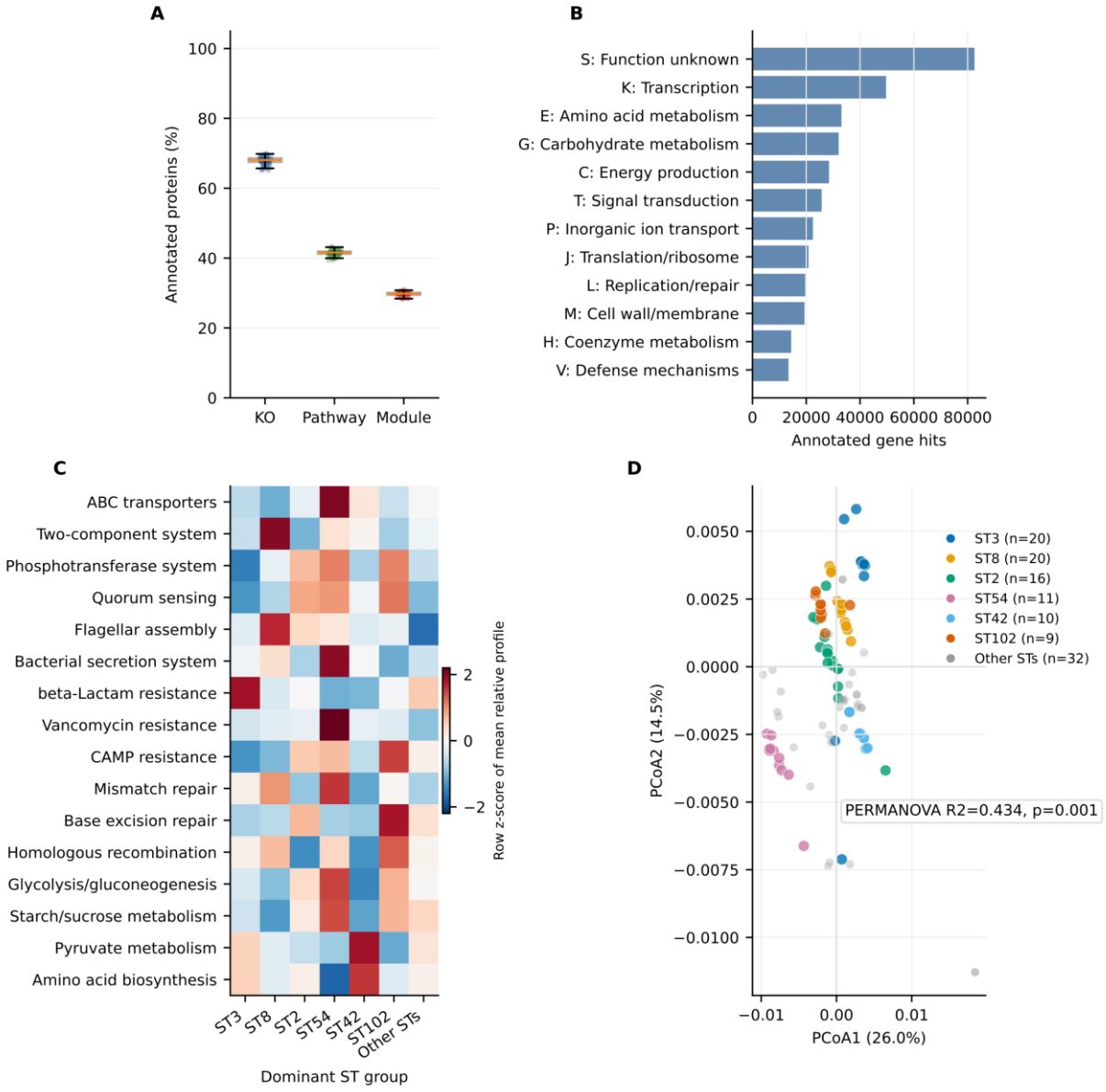

Figure 6. Standard functional annotation output for 118 *C. difficile* isolates. (A) Annotation coverage per isolate for KEGG Orthology, KEGG pathway and KEGG module assignments. Boxplots show the distribution across isolates, with individual isolate values overlaid. (B) Dominant COG functional categories ranked by annotated gene hits. (C) Selected KEGG pathway profile shifts by dominant ST group, shown as row z-scores of mean relative profiles. (D) Bray-Curtis PCoA of relative KEGG pathway profiles coloured by dominant ST group. PERMANOVA showed significant functional structuring by the dominant ST group (R^2 = 0.434, p = 0.001).

The dominant COG categories were function unknown (S; 82,551 gene hits), transcription (K; 49,668), amino acid metabolism (E; 33,080), carbohydrate metabolism (G; 32,046), energy production (C; 28,450) and signal transduction (T; 25,718) (Figure 6B). This profile indicates a conserved functional backbone but also shows that a large fraction of annotated gene content remains uncharacterized (19.9%). Comparison with the reference

strain CD630 revealed broadly conserved functional repertoires (Figure S2). Most central metabolic pathways, including glycolysis, pyruvate metabolism, butanoate metabolism, ABC transporters, and two-component systems, exhibited annotation burdens close to reference levels. These findings suggest that extensive accessory-genome diversity is superimposed on a relatively conserved functional backbone.

Selected KEGG pathway profiles varied by dominant ST group (Figure 6C). ST54 showed relatively higher profiles for ABC transporters, bacterial secretion system, vancomycin-resistance annotation category and mismatch repair, while ST42 showed a relative shift toward pyruvate metabolism and amino acid biosynthesis. These differences should be interpreted as annotation-profile shifts rather than direct evidence of phenotype, pathway activity or resistance.

Bray-Curtis PCoA of relative KEGG pathway profiles supported ST-associated structure (Figure 6D). PERMANOVA showed that dominant ST grouping explained 43.4% of Bray-Curtis variation in functional profiles ($R^2 = 0.434$, $p = 0.001$), with ST54 forming the clearest visually separated group. These results indicate that predicted functional repertoires are partly organised by lineage, but they require phenotypic or expression-level validation.

4. Discussion

This study places a *C. difficile* genome collection from Hong Kong, China, within a broader body of work showing that the species evolves through both clonal expansion and extensive genome remodelling. Early comparative analyses described *C. difficile* as a mosaic organism with abundant mobile DNA, and subsequent phylogenomic studies showed that clinically important lineages can spread over large geographic scales while continuing to accumulate recombination and accessory variation [6, 46]. Our results are consistent with this framework. The recombination-filtered tree retained clear lineage structure, but the pangenome, accessory-distance and MGE-context analyses showed that the core tree alone does not capture the full genomic diversity of the population.

The open pangenome observed here is therefore not an isolated feature of this dataset. Previous pangenome and systems-biology analyses have shown that *C. difficile* contains a large variable gene pool with discriminatory functional information across strains and lineages [47–49]. We identified 9,140 gene clusters, with most belonging to the cloud genome, and the Heaps' law exponent indicated that the *C. difficile* pangenome remains open. This demonstrates the high genetic diversity of the Hong Kong *C. difficile* population. Importantly, surveillance based only on core-genome relatedness may overlook differences in accessory genes that contribute to antimicrobial resistance, host adaptation, metabolic capacity, and environmental persistence.

The moderate association between accessory distance and recombination-filtered SNP distance helps refine how this population should be interpreted. Transmission and genomic epidemiology studies have established the value of core-genome SNPs for measuring recent relatedness, but they also show that *C. difficile* diversity cannot be reduced to a single distance measure [50–52]. In our analysis, accessory distance increased significantly with SNP distance, indicating that gene content is partly lineage-structured. At the same time, the broad scatter and the highlighted close isolate pairs show that recent relatives can differ in accessory content. This pattern supports a practical surveillance gap; future analyses should report both core-genome relatedness and accessory or mobile-element differences when assessing local clusters [46, 49].

A major finding of this study was the extensive association between antimicrobial-resistance loci and MGE signatures. The Sankey analysis revealed that resistance determinants were linked to multiple MGE categories, including *Tn916*-like transposons, IS*120*/IS*3*-family insertion sequences, IS*256*-family insertion sequences and recombinase/integrase-associated regions. These observations are consistent with the highly mobile nature of the *C. difficile* genome, in which approximately 11–30% of genomic content may be composed of MGE that contribute substantially to genome evolution, adaptation and AMR dissemination [53].

The broad distribution of MGE-associated loci across several STs suggests that horizontal gene transfer contributes to accessory-genome diversification throughout the population rather than within a single dominant lineage. This interpretation is consistent with previous studies showing that the *C. difficile* genome contains numerous conjugative transposons, insertion sequences and other mobile elements capable of mediating gene acquisition and genomic rearrangements [14, 54, 55].

Among the resistance determinants identified, *erm(B)* and *tet(M)* showed extensive connectivity to MGE-associated signals. This finding is concordant with previous genomic studies that identified these genes among the most prevalent acquired resistance determinants in the global *C. difficile* population. *erm(B)* is a major mediator of macrolide-lincosamide-streptogramin B resistance, whereas *tet(M)* is a well-established determinant of tetracycline resistance. The association between *tet(M)* and *Tn916*-like transposons observed in the present study is particularly notable. *Tn916*-family elements are among the best-characterised conjugative transposons in *C. difficile* and are recognised as important vehicles for the dissemination of tetracycline resistance genes. Previous investigations demonstrated that *Tn916* commonly carries *tet(M)* and contributes to resistance-gene transfer both within *C. difficile* and between bacterial species [56]. Similarly, previous analyses of *C. difficile* mobilomes have shown that *erm(B)* is frequently associated with transposons such as *Tn6194* and related conjugative elements, supporting the interpretation that horizontal transfer contributes to the widespread distribution of MLSB-resistance determinants [57, 58].

Interestingly, the most prevalent mobility signals in the present collection were IS*120*/IS*3*-family and IS*256*-family insertion sequences. These insertion-sequence families were associated with more isolates and loci than any other MGE category identified in the analysis. Insertion sequences are known drivers of bacterial genome evolution because they mediate transposition, genome rearrangement, gene disruption and mobilisation of adjacent DNA regions [59, 60]. Their widespread distribution across multiple STs in the present study suggests that local genome restructuring may be a major contributor to accessory-genome diversification in the Hong Kong *C. difficile* population. The prominence of IS*256* is particularly noteworthy. A large-scale analysis of over 1,101 *C. difficile* genomes identified IS*256* among the most frequently detected MGE and reported associations between IS*256*-containing elements and aminoglycoside-resistance genes [57].

This observation is consistent with our Sankey analysis, in which the aminoglycoside-resistance determinant *aac(6')-Ie/aph(2'')-Ia* exhibited strong connectivity to IS-associated mobility signals, supporting its association with MGE. Although aminoglycosides are not used for the treatment of *C. difficile* infection (CDI) and these genes are unlikely to confer a direct selective advantage during CDI, their presence may have broader ecological and clinical significance. The detection of mobile aminoglycoside-resistance determinants within *C. difficile* raises concerns regarding its potential role as a reservoir of transferable resistance genes within the intestinal microbiome [57], a setting in which antibiotic exposure can perturb microbial homeostasis and favour persistence of antibiotic-resistant opportunists [61, 62]. This is particularly important because aminoglycoside-modifying enzymes such as *aac(6')-Ie/aph(2'')-Ia* are well-recognised resistance determinants in major healthcare-associated pathogens, including *Enterococcus faecium*, *Staphylococcus aureus*, *Klebsiella pneumoniae*, *Acinetobacter baumannii*, *Pseudomonas aeruginosa* and *Enterobacter* spp. [63]. Collectively, these organisms constitute the ESKAPE group, which is responsible for a substantial proportion of multidrug-resistant healthcare-associated infections worldwide [64]. Clinical studies of *Acinetobacter baumannii* further illustrate how nosocomial pathogens can combine AMR with traits such as biofilm formation, strengthening the rationale for interpreting resistance genes in relation to ecological persistence as well as treatment failure [65]. The coexistence of *Tn916*-like transposons, IS families and recombinase/integrase-associated regions in our dataset suggests that multiple mobility mechanisms may facilitate the persistence and movement of resistance genes within the gut microbial community. While direct interspecies transfer cannot be inferred from the present data, the genomic context observed here supports the hypothesis that *C. difficile* may contribute to the wider gut resistome and potentially serve as a source of resistance determinants that can be mobilised into clinically significant pathogens.

Despite the widespread occurrence of MGE-associated loci, a substantial proportion of resistance loci lacked detectable mobile-element signatures. Several explanations may account for this observation. Resistance genes may have become stably integrated into the

chromosome following historical acquisition events [66], or the mobile-element boundaries may be incompletely resolved because the present study relied on short-read draft assemblies [67]. Furthermore, some mobile elements may not be represented in the reference databases used for screening. Therefore, the absence of a detectable MGE signal should not be interpreted as evidence of exclusively vertical inheritance.

Functional annotation revealed that the 118 *C. difficile* isolates shared a broadly conserved functional backbone, but KEGG pathway profiles still showed measurable ST-associated structuring. This is consistent with systems-level pangenome work showing that *C. difficile* gene-content diversity contains strain-discriminatory functional information, even when core metabolic functions remain conserved [47]. The dominance of "function unknown" among COG categories also reflects a common limitation of bacterial genome annotation, where many predicted proteins remain poorly characterised [68, 69]. Therefore, the COG result should be interpreted mainly as a broad functional overview rather than a direct biological mechanism.

The ST-associated KEGG shifts, especially in transport, two-component systems, secretion, resistance-related pathways, motility and DNA repair categories, suggest that lineage-level differences may involve environmental sensing, nutrient acquisition, stress adaptation and genome maintenance. These functional categories are biologically plausible in *C. difficile* because adaptation to the gut environment, antimicrobial exposure and host-associated stress are central to CDI ecology. ST54 is particularly interesting because it has been reported as a common or dominant ST in several Chinese clinical collections, including tertiary-hospital, military-hospital and community-onset CDI studies [70–72]. However, those studies mainly focused on molecular epidemiology, toxin profiles and antimicrobial susceptibility rather than ST-specific KEGG functional signatures. Therefore, the ST54-associated transporter, secretion, resistance-category and repair-associated profile observed here should be interpreted as a hypothesis-generating functional pattern rather than evidence of a confirmed ST54 phenotype. These results suggest that functional annotation can provide an additional layer of lineage discrimination beyond MLST and phylogeny. However, because KEGG and COG profiles are annotation-derived, the observed ST-associated structure requires validation using isolate-level KO inspection, long-read assemblies, transcriptomics, metabolomics or targeted phenotypic assays.

Much of the foundational *C. difficile* genomic literature has focused on global epidemic lineages, European and North American transmission settings, or specific high-profile lineages such as RT027 and ST11 [4, 11, 50]. Regional datasets from Asia are essential because local antimicrobial use, healthcare networks, animal reservoirs and lineage composition can shape the accessory genome differently. The present analysis does not claim to define all Hong Kong *C. difficile* diversity, but it provides a structured regional baseline

linking phylogeny, pangenome openness, AMR context, toxin-locus interpretation and functional annotation in the same isolate set.

The study has several limitations. Clinical metadata, patient movement, ward overlap and antimicrobial susceptibility data were unavailable [73], so close genomic relationships cannot be interpreted as direct transmission or phenotypic resistance. The analyses were performed on short-read assemblies, which may fragment plasmids, prophages, conjugative transposons and toxin loci. Local MGE-neighbourhood evidence should therefore be interpreted as proximity-based genomic support rather than proof of transferability. Rare STs were represented by small numbers of isolates, limiting lineage-specific inference outside the dominant STs.

Future work should combine long-read sequencing, read-linkage or PCR validation of candidate AMR-MGE modules, antimicrobial susceptibility testing, toxin phenotyping, and prospective clinical metadata. These additions would allow candidate resistance modules, toxin-locus structures, close genomic groups, and ST-associated functional signatures to be tested as biological and epidemiological hypotheses. Despite these limitations, the present study provides a regional genomic baseline for *C. difficile* in Hong Kong and supports surveillance frameworks that integrate lineage, core-genome relatedness, accessory-gene content, AMR detection, and MGE-context evidence.

5. Conclusions

This study provides a genome-resolved overview of *C. difficile* diversity in Hong Kong, China, revealing a population structured by distinct phylogenetic lineages alongside substantial accessory genome variability. The open pangenome and increasing accessory-genome divergence with core-genome SNP distance indicate ongoing genomic diversification within the population. Antimicrobial-resistance determinants were heterogeneously distributed and frequently associated with mobile-element contexts, whereas major toxin loci remained broadly conserved. Functional annotation identified lineage-associated differences in predicted gene content, including a distinctive ST54 functional profile that warrants further investigation. Collectively, these findings establish an important regional genomic surveillance baseline and demonstrate the value of integrating phylogenomics, pangenome analysis, mobile-element context and functional annotation to investigate the evolution and dissemination of clinically relevant *C. difficile* lineages.

Supplementary Materials: Table S1: Galaxy Tool version and parameters. Figure S1: Pangenome alignment of 118 *Clostridioides difficile* isolates; Figure S2: CD630 reference-normalised eggNOG functional comparison.

Author Contributions: Conceptualisation, LT-HT, SHW and L-HL; methodology, LT-HT; formal analysis, LT-HT, K-OC, L-HL; investigation, LT-HT, K-OC, JW-FL; resources, MI, SHW, L-HL; writing, review and editing, LT-HT, JW-FL, K-OC, K-GC, RWYN, WSL, MI, SHW and L-HL.

Funding: This work was partially supported by the Commissioned Health and Medical Research Fund (HMRF), Food and Health Bureau of Hong Kong, Special Administrative Region, Grant no. CID-CUHK-C.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. McDonald LC, Gerding DN, Johnson S, *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). Clinical Infectious Diseases 2018; 66(7): e1–e48. doi: 10.1093/cid/cix1085.
2. Ho J, Dai RZW, Kwong TNY, *et al*. Disease Burden of Clostridium difficile Infections in Adults, Hong Kong, China, 2006-2014. Emerging Infectious Diseases 2017; 23(10): 1671–1679. doi: 10.3201/eid2310.170797.
3. Collins DA, Sohn KM, Wu Y, *et al*. Clostridioides difficile infection in the Asia-Pacific region. Emerging Microbes & Infections 2020; 9(1): 42–52. doi: 10.1080/22221751.2019.1702480.
4. He M, Miyajima F, Roberts P, *et al*. Emergence and global spread of epidemic healthcare-associated Clostridium difficile. Nature Genetics 2013; 45(1): 109–13. doi: 10.1038/ng.2478.
5. Chong KJ, Feng H, Letchumanan V, *et al*. Tackling Microbial Resistance and Emerging Pathogens with Next-Generation Antibiotics. Progress In Microbes and Molecular Biology 2024; 7(1): a0000447. doi: 10.36877/pmmb.a0000447.
6. Beebe MA and Sorg JA. An update on *Clostridioides difficile* population structure and genomics. Clinical Microbiology Reviews 2026; 39(2): e00296–25. doi: doi:10.1128/cmr.00296-25.
7. Dingle KE, Elliott B, Robinson E, *et al*. Evolutionary history of the Clostridium difficile pathogenicity locus. Genome Biology and Evolution 2014; 6(1): 36–52. doi: 10.1093/gbe/evt204.
8. Sebaihia M, Wren BW, Mullany P, *et al*. The multidrug-resistant human pathogen Clostridium difficile has a highly mobile, mosaic genome. Nature Genetics 2006; 38(7): 779–86. doi: 10.1038/ng1830.
9. Griffiths D, Fawley W, Kachrimanidou M, *et al*. Multilocus sequence typing of Clostridium difficile. Journal of Clinical Microbiology 2010; 48(3): 770–8. doi: 10.1128/jcm.01796-09.
10. Jolley KA, Bray JE, and Maiden MCJ. Open-access bacterial population genomics: BIGSdb software, the PubMLST.org website and their applications. Wellcome Open Research 2018; 3: 124. doi: 10.12688/wellcomeopenres.14826.1.
11. Eyre DW, Cule ML, Wilson DJ, *et al*. Diverse sources of C. difficile infection identified on whole-genome sequencing. New England Journal of Medicine 2013; 369(13): 1195–205. doi: 10.1056/NEJMoa1216064.
12. Didelot X, Eyre DW, Cule M, *et al*. Microevolutionary analysis of Clostridium difficile genomes to investigate transmission. Genome Biology 2012; 13(12): R118. doi: 10.1186/gb-2012-13-12-r118.
13. Roberts AP and Mullany P. Tn916-like genetic elements: a diverse group of modular mobile elements conferring antibiotic resistance. FEMS Microbiology Reviews 2011; 35(5): 856–71. doi: 10.1111/j.1574-6976.2011.00283.x.
14. Spigaglia P. Recent advances in the understanding of antibiotic resistance in Clostridium difficile infection. Therapeutic Advances in Infectious Disease 2016; 3(1): 23–42. doi: 10.1177/2049936115622891.
15. Ortiz-Flores CG, Romero-Rodríguez A, Villanueva-Enríquez R, *et al*. Comparative Genomic Analysis of Clostridioides difficile Strains in Mexico: Insights into Virulence and Resistance. Access Microbiology 2026. doi: 10.1099/acmi.0.001132.v2.
16. Cantalapiedra CP, Hernández-Plaza A, Letunic I, *et al*. eggNOG-mapper v2: Functional Annotation, Orthology Assignments, and Domain Prediction at the Metagenomic Scale. Molecular Biology and Evolution 2021; 38(12): 5825–5829. doi: 10.1093/molbev/msab293.

17. Huerta-Cepas J, Szklarczyk D, Heller D, *et al*. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. Nucleic Acids Research 2019; 47(D1): D309–D314. doi: 10.1093/nar/gky1085.
18. Page AJ, Cummins CA, Hunt M, *et al*. Roary: rapid large-scale prokaryote pan genome analysis. Bioinformatics 2015; 31(22): 3691–3693. doi: 10.1093/bioinformatics/btv421.
19. Camargo AP, Roux S, Schulz F, *et al*. Identification of mobile genetic elements with geNomad. Nature Biotechnology 2024; 42(8): 1303–1312. doi: 10.1038/s41587-023-01953-y.
20. Xie Z and Tang H. ISEScan: automated identification of insertion sequence elements in prokaryotic genomes. Bioinformatics 2017; 33(21): 3340–3347. doi: 10.1093/bioinformatics/btx433.
21. Kong KY, Kwong TNY, Chan H, *et al*. Biological characteristics associated with virulence in Clostridioides difficile ribotype 002 in Hong Kong. Emerging Microbes & Infections 2020; 9(1): 631–638. doi: 10.1080/22221751.2020.1739564.
22. Wong SH, Ip M, Hawkey PM, *et al*. High morbidity and mortality of Clostridium difficile infection and its associations with ribotype 002 in Hong Kong. Journal of Infection 2016; 73(2): 115–122. doi: https://doi.org/10.1016/j.jinf.2016.05.010.
23. Chow VCY, Kwong TNY, So EWM, *et al*. Surveillance of antibiotic resistance among common Clostridium difficile ribotypes in Hong Kong. Scientific Reports 2017; 7(1): 17218. doi: 10.1038/s41598-017-17523-7.
24. Goh SY, Chua K-O, Khan SA, *et al*. Genomic and Resistome Analyses of Elizabethkingia anophelis Strain B2D isolated from Dental Plaque of Patient. Progress In Microbes and Molecular Biology 2024; 7(1): a0000440. doi: 10.36877/pmmb.a0000440.
25. Mansfield MJ, Tremblay BJ, Zeng J, *et al*. Phylogenomics of 8,839 Clostridioides difficile genomes reveals recombination-driven evolution and diversification of toxin A and B. PLoS Pathogens 2020; 16(12): e1009181. doi: 10.1371/journal.ppat.1009181.
26. The Galaxy Community. The Galaxy platform for accessible, reproducible, and collaborative data analyses: 2024 update. Nucleic Acids Research 2024; 52(W1): W83–W94. doi: 10.1093/nar/gkae410.
27. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics 2014; 30(14): 2068–9. doi: 10.1093/bioinformatics/btu153.
28. Seemann T. Snippy: rapid haploid variant calling and core SNP phylogeny. GitHub. Available at: github. com/tseemann/snippy 2015.
29. Page AJ, Taylor B, Delaney AJ, *et al*. SNP-sites: rapid efficient extraction of SNPs from multi-FASTA alignments. Microbial Genomics 2016; 2(4). doi: https://doi.org/10.1099/mgen.0.000056.
30. Croucher NJ, Page AJ, Connor TR, *et al*. Rapid phylogenetic analysis of large samples of recombinant bacterial whole genome sequences using Gubbins. Nucleic Acids Research 2015; 43(3): e15–e15. doi: 10.1093/nar/gku1196.
31. Minh BQ, Schmidt HA, Chernomor O, *et al*. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 2020; 37(5): 1530–1534. doi: 10.1093/molbev/msaa015.
32. Kalyaanamoorthy S, Minh BQ, Wong TKF, *et al*. ModelFinder: fast model selection for accurate phylogenetic estimates. Nature Methods 2017; 14(6): 587–589. doi: 10.1038/nmeth.4285.
33. Letunic I and Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. Nucleic Acids Research 2019; 47(W1): W256–w259. doi: 10.1093/nar/gkz239.

34. Feldgarden M, Brover V, Gonzalez-Escalona N, *et al*. AMRFinderPlus and the Reference Gene Catalog facilitate examination of the genomic links among antimicrobial resistance, stress response, and virulence. Scientific Reports 2021; 11(1): 12728. doi: 10.1038/s41598-021-91456-0.
35. Seemann T. ABRicate: mass screening of contigs for antiobiotic resistance genes. 2016. URL https://github.com/tseemann/abricate 2024.
36. Alcock BP, Raphenya AR, Lau TTY, *et al*. CARD 2020: antibiotic resistome surveillance with the comprehensive antibiotic resistance database. Nucleic Acids Research 2020; 48(D1): D517–D525. doi: 10.1093/nar/gkz935.
37. Gupta SK, Padmanabhan BR, Diene SM, *et al*. ARG-ANNOT, a New Bioinformatic Tool To Discover Antibiotic Resistance Genes in Bacterial Genomes. Antimicrobial Agents and Chemotherapy 2014; 58(1): 212–220. doi: doi:10.1128/aac.01310-13.
38. Zankari E, Hasman H, Cosentino S, *et al*. Identification of acquired antimicrobial resistance genes. Journal of Antimicrobial Chemotherapy 2012; 67(11): 2640–2644. doi: 10.1093/jac/dks261.
39. Chen L, Zheng D, Liu B, *et al*. VFDB 2016: hierarchical and refined dataset for big data analysis—10 years on. Nucleic Acids Research 2016; 44(D1): D694–D697. doi: 10.1093/nar/gkv1239.
40. Ben Maamar S, Glawe AJ, Brown TK, *et al*. Mobilizable antibiotic resistance genes are present in dust microbial communities. PLoS Pathogens 2020; 16(1): e1008211. doi: 10.1371/journal.ppat.1008211.
41. Sun J, Liao X-P, D'Souza AW, *et al*. Environmental remodeling of human gut microbiota and antibiotic resistome in livestock farms. Nature communications 2020; 11(1): 1427. doi: 10.1038/s41467-020-15222-y.
42. Shi Z, Liu Q, Zhou M, *et al*. Persistent Risks in the Effluents of Wastewater Treatment Plants: Mobile Genetic Elements and Viral-Mediated Dissemination of Pathogenic Antibiotic-Resistant Bacteria. Environmental Science & Technology 2025; 59(43): 23374–23385. doi: 10.1021/acs.est.5c08352.
43. Carattoli A, Zankari E, García-Fernández A, *et al*. *In Silico* Detection and Typing of Plasmids using PlasmidFinder and Plasmid Multilocus Sequence Typing. Antimicrobial Agents and Chemotherapy 2014; 58(7): 3895–3903. doi: doi:10.1128/aac.02412-14.
44. Buchfink B, Reuter K, and Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. Nature Methods 2021; 18(4): 366–368. doi: 10.1038/s41592-021-01101-x.
45. Kanehisa M, Furumichi M, Sato Y, *et al*. KEGG: integrating viruses and cellular organisms. Nucleic Acids Research 2021; 49(D1): D545–D551. doi: 10.1093/nar/gkaa970.
46. Sholeh M, Beig M, Kouhsari E, *et al*. Global insights into the genome dynamics of Clostridioides difficile associated with antimicrobial resistance, virulence, and genomic adaptations among clonal lineages. Frontiers in Cellular and Infection Microbiology 2025; Volume 14 - 2024. doi: 10.3389/fcimb.2024.1493225.
47. Norsigian CJ, Danhof HA, Brand CK, *et al*. Systems biology approach to functionally assess the *Clostridioides difficile* pangenome reveals genetic diversity with discriminatory power. Proceedings of the National Academy of Sciences 2022; 119(18): e2119396119. doi: doi:10.1073/pnas.2119396119.
48. Golchha NC, Nighojkar A, and Nighojkar S. Redefining genomic view of Clostridioides difficile through pangenome analysis and identification of drug targets from its core genome. Drug Target Insights 2022; 16: 17–24. doi: 10.33393/dti.2022.2469.
49. Dost I, Abdel-Glil M, Persson S, *et al*. Genomic study of European Clostridioides difficile ribotype 002/sequence type 8. Microbial Genomics 2024; 10(7). doi: https://doi.org/10.1099/mgen.0.001270.
50. Knight DR, Kullin B, Androga GO, *et al*. Evolutionary and Genomic Insights into Clostridioides difficile Sequence Type 11: a Diverse Zoonotic and Antimicrobial-Resistant Lineage of Global One Health Importance. mBio 2019; 10(2): 10.1128/mbio.00446–19. doi: doi:10.1128/mbio.00446-19.

51. Sundermann AJ, Mills EG, Rangachar Srinivasa V, *et al*. Enhanced discrimination of Clostridioides difficile transmission using whole-genome sequencing and in silico multi-locus variable number tandem repeat analysis. Microbiology Spectrum 2025; 13(12): e0328724. doi: 10.1128/spectrum.03287-24.
52. Newcomer EP, Fishbein SRS, Zhang K, *et al*. Genomic surveillance of Clostridioides difficile transmission and virulence in a healthcare setting. mBio 2024; 15(3): e0330023. doi: 10.1128/mbio.03300-23.
53. Mullany P, Allan E, and Roberts AP. Mobile genetic elements in Clostridium difficile and their role in genome function. Research in Microbiology 2015; 166(4): 361–7. doi: 10.1016/j.resmic.2014.12.005.
54. Brouwer MSM, Warburton PJ, Roberts AP, *et al*. Genetic Organisation, Mobility and Predicted Functions of Genes on Integrated, Mobile Genetic Elements in Sequenced Strains of Clostridium difficile. PloS One 2011; 6(8): e23014. doi: 10.1371/journal.pone.0023014.
55. Blau K and Gallert C. Mobilome of Environmental Isolates of Clostridioides difficile. Antibiotics 2025; 14(7): 678. doi: https://doi.org/10.3390/antibiotics14070678.
56. Wu Y, Liu C, Li W-G, *et al*. Independent Microevolution Mediated by Mobile Genetic Elements of Individual Clostridium difficile Isolates from Clade 4 Revealed by Whole-Genome Sequencing. mSystems 2019; 4(2): 10.1128/msystems.00252–18. doi: doi:10.1128/msystems.00252-18.
57. Kartalidis P, Skoulakis A, Tsilipounidaki K, *et al*. Clostridioides difficile as a Dynamic Vehicle for the Dissemination of Antimicrobial-Resistance Determinants: Review and In Silico Analysis. Microorganisms 2021; 9(7). doi: 10.3390/microorganisms9071383.
58. O'Grady K, Hong S, Putsathit P, *et al*. Defining the phylogenetics and resistome of the major Clostridioides difficile ribotypes circulating in Australia. Microbial Genomics 2024; 10(5). doi: 10.1099/mgen.0.001232.
59. Siguier P, Gourbeyre E, and Chandler M. Bacterial insertion sequences: their genomic impact and diversity. FEMS Microbiol Rev 2014; 38(5): 865–91. doi: 10.1111/1574-6976.12067.
60. Ngan WY, Parab L, Bertels F, *et al*. A more significant role for insertion sequences in large-scale rearrangements in bacterial genomes. mBio 2025; 16(1): e03052–24. doi: doi:10.1128/mbio.03052-24.
61. Thye AY-K, Kumari Y, Chan K-G, *et al*. In Vitro Antagonistic Activity of Probiotic Strains against Antibiotic-Resistant Klebsiella pneumoniae isolated from Preterm Infant Stools. Progress In Microbes and Molecular Biology 2025; 8(1): a0000477. doi: 10.36877/pmmb.a0000477.
62. Lau AWY, Tan LT-H, Ab Mutalib N-S, *et al*. The chemistry of gut microbiome in health and diseases. Progress In Microbes and Molecular Biology 2021; 4(1): a0000175. doi: 10.36877/pmmb.a0000175.
63. Ngoi ST, Chong CW, Ponnampalavanar SSLS, *et al*. Genetic mechanisms and correlated risk factors of antimicrobial-resistant ESKAPEE pathogens isolated in a tertiary hospital in Malaysia. Antimicrobial Resistance & Infection Control 2021; 10(1): 70. doi: 10.1186/s13756-021-00936-5.
64. Rizki LP, Murni IK, Aman AT, *et al*. Environmental Metagenomic Analysis of "ESKAPE" Pathogens in the Pediatric Intensive Care Unit of General Hospital Yogyakarta Indonesia. Progress In Microbes and Molecular Biology 2024; 7(1): a0000398. doi: 10.36877/pmmb.a0000398.
65. Ghimire U, Kandel R, Neupane M, *et al*. Biofilm Formation and blaOXA Genes Detection Among Acinetobacter baumannii from Clinical Isolates in a Tertiary Care Kirtipur Hospital, Nepal. Progress In Microbes and Molecular Biology 2021; 4(1): a0000245. doi: 10.36877/pmmb.a0000245.
66. Arnold BJ, Huang IT, and Hanage WP. Horizontal gene transfer and adaptive evolution in bacteria. Nat Rev Microbiol 2022; 20(4): 206–218. doi: 10.1038/s41579-021-00650-4.
67. Partridge SR, Kwong SM, Firth N, *et al*. Mobile Genetic Elements Associated with Antimicrobial Resistance. Clinical Microbiology Reviews 2018; 31(4). doi: 10.1128/cmr.00088-17.

68. Näpflin N, Schubert C, Malfertheiner L, *et al*. Gene-level analysis of core carbohydrate metabolism across the Enterobacteriaceae pan-genome. Communications Biology 2025; 8(1): 1241. doi: 10.1038/s42003-025-08640-5.
69. Gardoul M, Rached B, Mbarki A, *et al*. Comprehensive whole-genome analysis of Streptococcus infantarius strains from Moroccan farmhouse dairy products: Genomic insights into dairy adaptation, safety, and biotechnological potential. International Journal of Food Microbiology 2025; 442: 111358. doi: 10.1016/j.ijfoodmicro.2025.111358.
70. Chen YB, Gu SL, Wei ZQ, *et al*. Molecular epidemiology of Clostridium difficile in a tertiary hospital of China. Journal of Medical Microbiology 2014; 63(Pt 4): 562–569. doi: 10.1099/jmm.0.068668-0.
71. Wang R, Suo L, Chen HX, *et al*. Molecular epidemiology and antimicrobial susceptibility of Clostridium difficile isolated from the Chinese People's Liberation Army General Hospital in China. International Journal of Infectious Diseases 2018; 67: 86–91. doi: 10.1016/j.ijid.2017.07.010.
72. Jiang X, Bian J, Lv T, *et al*. Clinical characteristics of community-onset Clostridioides difficile infections at a tertiary hospital in mainland China: A fourteen-year (2010–2023) retrospective study. International Journal of Medical Microbiology 2024; 316: 151631. doi: 10.1016/j.ijmm.2024.151631.
73. Eyre DW, Laager M, Walker AS, *et al*. Probabilistic transmission models incorporating sequencing data for healthcare-associated Clostridioides difficile outperform heuristic rules and identify strain-specific differences in transmission. PLoS Computational Biology 2021; 17(1): e1008417. doi: 10.1371/journal.pcbi.1008417.